Exact Tail Probabilities and Percentiles of the Multinomial Maximum

Anirban DasGupta
Purdue University, USA

Abstract

The maximum cell frequency in a multinomial distribution is of current interest in several areas of probability and statistics. Different asymptotics apply for different rates of growth for the number of cells and the number of units. Statistical software for calculating the P-values or for calculating the percentiles is not presently available. Using the Poissonization theorem for multinomials, exact P-values and exact 95th and 99th percentiles are tabulated for a selection of values of the number of cells and units. The sparse multinomial case is included to some extent. Using the asymptotic extreme value theory, approximate formulas for percentiles are given for use outside of the range of the tables provided here.

1 Introduction

The maximum cell frequency in a multinomial distribution is of wide interest in cluster detection, data mining, goodness of fit, and in occupancy problems in probability. It also arises in sequential clinical trials, and in paranormal experiments. See some evidence in Diaconis and Graham (1981), Levin (1983), and Rukhin (2006). The multinomial maximum has also become important in the modern *large p small n* problems, where a small number of units are allocated among a large number of categories. See Hall and Titterington (1987), Koehler and Larntz (1980), Simonoff (1983), and Zelterman (1987) for treatment of sparse multinomial data. We provide some coverage of the sparse case in what follows. Examination of the exact P-values in the sparse case shows the need for much caution before concluding systematic departure from uniformity, because one should matter of factly expect a large number of empty cells, coupled with large values for the maximum cell count.

Most people do not have a very good intuition about what constitutes an extreme value for an extreme type statistic. If fifteen of fifty rolls of a die resulted in one particular face, we may suspect that the die has been manipulated. Actually, in that example, it is not so surprising to see some face come up fifteen times, when we actually compute the P-value. Asymptotic theory, both first order and higher order, for the maximum cell frequency in a multinomial distribution certainly exists; see Kolchin et al. (1978), and Barbour et al. (1992). These can be and are sometimes used to approximate P-values based on a maximum cell frequency, i.e., to approximate tail probabilities $P(\max\{f_1, f_2, \cdots, f_K\} \geq m)$ in the equiprobable case, where $f_1, f_2, \cdots, f_K$ denote the cell frequencies in a K-cell multinomial, and m is a given number. It is a classic result in probability that Poissonizing the total number of units in a multinomial problem renders the cell frequencies to become independent Poisson random variables. Precisely, if $N \sim \text{Poisson}(\lambda)$, and given $N = n, (f_1, f_2, \cdots, f_K)$ has a multinomial distribution with parameters $(n, p_1, p_2, \cdots, p_K)$, then unconditionally, $f_1, f_2, \cdots, f_K$ are independent and $f_i \sim \text{Poisson}(\lambda p_i)$. It follows that with any given fixed value n, and any given fixed set A in the K-dimensional Euclidean space $\mathcal{R}^K$, the multinomial probability that $(f_1, f_2, \cdots, f_K)$ belongs to A equals $n!c(n, \lambda)$, with $c(n, \lambda)$ being the coefficient of λ^n in the power series expansion of $e^\lambda P((X_1, X_2, \cdots, X_K) \in A)$, where now X_i are independent Poisson(λp_i). In the *equiprobable case*, i.e., when the p_i are all equal to $\frac{1}{K}$, this leads to the equality that $P(\max\{f_1, f_2, \cdots, f_K\} \geq x)$ is $\frac{n!}{K^n} \times$ The coefficient of λ^n in $(\sum_{j=0}^{x-1} \frac{\lambda^j}{j!})^K$. As a result, we can compute the P-value $P(\max\{f_1, f_2, \cdots, f_K\} \geq x)$ *exactly* whenever we can have a computer produce for us the coefficient of λ^n in the expansion of $(\sum_{j=0}^{x-1} \frac{\lambda^j}{j!})^K$. Common statistical software does not treat this problem, but it is possible to write a code on symbolic software to produce this coefficient.

The multinomial maximum being of enough interest right now, and with there being no common statistical software that computes these P-values for testing for uniformity, we thought that a table of the exact P-values, and moreover the 95th and the 99th percentiles would be useful. Obviously, necessarily any such table has to choose values of n and K. We generally limit ourselves to $n \leq 500$ and to $K \leq 12$. The values $K = 7, 12$ are of some special interest, being the number of days in a week, and the number of months in a year. Likewise, $K = 30$ and $K = 365$ would also be of some special importance; but we did not go that far here. One final remark is that the 95th percentile was chosen to be the first value x such that $P(\max\{f_1, f_2, \cdots, f_K\} \leq x) \geq .95$, *unless* the previous value gave a probability extremely close to .95. Similar comments apply about the 99th percentile.

The 95th and the 99th percentiles are tabulated first, and then more elaborate tables of the actual P-values are given. These P-values are all exact; no simulations or approximations are involved.

2 Approximate Formulas for Practical Use

It is obviously impossible to produce tables of percentiles for all or even many combinations of n and K. On the other hand, no readymade statistical software for accurate calculation of the percentiles or the tail probabilities seems to be currently available. Therefore, approximate formulas for specific percentiles of the multinomial maximum have practical value. We provide here such approximate formulas when n and K are both large. The formulas are based on asymptotic theory for the multinomial maximum. The asymptotics for the maximum frequency are known to depend on the relative growth of n and K. We provide approximate formulas for percentiles when n and K are both large, and moreover n is substantially larger than K. It is in this case that the formulas are the most trustable and the easiest to derive.

If K is substantially larger than n, then eventually the distribution of the multinomial maximum becomes a two-point distribution. *This is why we see the rapid drop in the P-values in our tables under the sparse case.* It is not very useful to provide something like a 95th percentile when the distribution is essentially two valued.

If n and K are comparable, in the sense that $\frac{n}{K \log K}$ is not much larger than one, the distribution of the multinomial maximum becomes more dispersed and asymptotically it is supported on a countable set of integers, lower bounded by a suitable integer. However, identifying this lower bound in a given problem so as to make the approximation accurate is problematic. It involves a case by case trial and error, defeating the entire purpose of a theoretical approximate formula. This is why we limit ourselves to the case when n is substantially larger than K.

The approximate formula is based on the following theorem (see Kolchin et al. (1978), pp 96, Theorem 3, after correcting the typographical errors in the statement). In the theorem, f_{max} denotes $\max\{f_1, f_2, \cdots, f_K\}$.

Theorem Let $(f_1, f_2, \cdots, f_K) \sim \text{Mult}(n, \frac{1}{K}, \cdots, \frac{1}{K})$. Suppose $n, K = K(n) \to \infty$ such that $\frac{n}{K \log K} \to \infty$. Let

$$\mu = \mu(n) = \frac{n}{K}; w = w(n) = \frac{\log K - \frac{1}{2} \log \log K}{\mu};$$

$\epsilon = \epsilon(n)$ the unique positive root of the equation

$$(1 + \epsilon) \log(1 + \epsilon) - \epsilon = w.$$

Then,

$$P(\frac{f_{max} - \mu(1+\epsilon)}{\sqrt{\frac{n}{2K\log K}}} + \frac{1}{2}\log 4\pi \leq z) \to e^{-e^{-z}},$$

for all real z.

This result leads to simple enough approximate formulas for percentiles of f_{max}. A first order approximation to ϵ in the statement of the theorem is $\epsilon = \sqrt{\frac{K\log K}{n}}$, on writing $\log(1+\epsilon) \approx \epsilon$. Inverting the CDF $Q(z) = e^{-e^{-z}}$, for any $\alpha, 0 < \alpha < 1$, the $100(1-\alpha)\%$ percentile of Q is $-\log\log\frac{1}{1-\alpha}$. A few lines of algebra then produces the approximate formula for the $100(1-\alpha)\%$ percentile F_α of f_{max} as:

$$F_\alpha \approx \frac{n}{K} + \sqrt{\frac{n\log K}{K}} - (\log\log\frac{1}{1-\alpha} + 1.266)\sqrt{\frac{n}{2K\log K}},$$

provided, n, K, and $\frac{n}{K\log K}$ are each large. In the above, we have used the decimal value 1.266 for $\frac{1}{2}\log 4\pi$.

In particular, approximate 95th and 99th percentiles of f_{max} are:

$$F_{.05} \approx \frac{n}{K} + \sqrt{\frac{n\log K}{K}} + 1.205\sqrt{\frac{n}{K\log K}};$$

$$F_{.01} \approx \frac{n}{K} + \sqrt{\frac{n\log K}{K}} + 2.358\sqrt{\frac{n}{K\log K}}.$$

As a trial, if we use these approximate formulas when $K = 12, n = 400$, we get $F_{.05} \approx 47$, and $F_{.01} \approx 51$, while the true values are 50 and 53, respectively (see table on pp 8). As another trial case, with $K = 10, n = 500$, we get $F_{.05} \approx 66$, and $F_{.01} \approx 72$, while the true values are 69 and 73, respectively. The approximations seem quite good even when K is not that large.

3 Table of Percentiles

Table of 95th and 99th Percentiles of $\max\{f_1, f_2, \cdots, f_K\}$

n	95th Percentile	99th Percentile
	$K = 3$	
10	7	8
15	10	11
25	14	16
40	21	23
50	25	27
75	35	37
100	44	47
150	63	67
200	82	86
250	100	105
	$K = 4$	
10	7	8
25	12	14
40	17	19
50	20	22
75	28	21
100	36	38
150	51	54
200	65	69
250	79	83
	$K = 5$	
10	6	7
25	10	11
50	18	20
75	24	26
100	31	33
150	43	46
200	54	58
250	66	70
300	77	81

Table of 95th and 99th Percentiles of $\max\{f_1, f_2, \cdots, f_K\}$

n	95th Percentile	99th Percentile
	$K = 6$	
10	6	7
25	10	11
50	16	18
100	27	29
150	37	40
200	47	50
250	57	61
300	67	70
	$K = 7$	
25	9	10
50	14	16
100	24	26
150	33	36
200	42	45
250	51	54
300	59	63
400	76	80
	$K = 10$	
50	12	13
100	19	21
200	32	35
300	45	48
400	57	60
500	69	73

Table of 95th and 99th Percentiles of $\max\{f_1, f_2, \cdots, f_K\}$

n	95th Percentile	99th Percentile
	$K = 12$	
50	11	12
100	17	19
200	29	31
300	39	42
400	50	53
500	60	63

4 Table of Tail Probabilities

$P(\max\{f_1, f_2, \cdots, f_K\} \geq x)$ $(K = 5)$

x	$n = 25$	$n = 50$	$n = 60$	$n = 75$	$n = 100$
8	.5100	1	1	1	1
9	.2311	1	1	1	1
10	.0866	1	1	1	1
11	.0278	.9995	1	1	1
12	.0077	.9497	1	1	1
13	.0018	.7646	.9996	1	1
14	.0004	.5119	.9631	1	1
15	.00007	.2960	.8143	1	1
16	.00001	.1530	.5875	.9998	1
17	1.33×10^{-6}	.0721	.3707	.9750	1
18	0	.0313	.2106	.8641	1
19	0	.0126	.1099	.6733	1
20	0	.0047	.0533	.4664	1
21	0	.0016	.0241	.2936	.9999
22	0	.0005	.0102	.1711	.9851
23	0	.00015	.0041	.0933	.9119
24	0	.00004	.0015	.0480	.7680
25	0	.00001	.0005	.0233	.5878
26	0	2.46×10^{-6}	.00017	.0107	.4141
27	0	0	.00005	.0047	.2724
28	0	0	.00002	.0019	.1690
29	0	0	4.07×10^{-6}	.00076	.0997
30	0	0	1.03×10^{-6}	.00028	.0562
31	0	0	0	.0001	.0303
32	0	0	0	.00003	.0156
33	0	0	0	.00001	.0078
34	0	0	0	3.22×10^{-6}	.0037
35	0	0	0	0	.0017
36	0	0	0	0	.00074
37	0	0	0	0	.00031
38	0	0	0	0	.00012
39	0	0	0	0	.00005
40	0	0	0	0	.00002

$P(\max\{f_1, f_2, \cdots, f_K\} \geq x)$ $(K = 6)$

x	$n = 30$	$n = 50$	$n = 100$	$n = 120$	$n = 150$
8	.6014	1	1	1	1
9	.2942	1	1	1	1
10	.1176	.9888	1	1	1
11	.0404	.8663	1	1	1
12	.0122	.6122	1	1	1
13	.0032	.3578	1	1	1
14	.00076	.1816	1	1	1
15	.00016	.0827	1	1	1
16	.00003	.0344	1	1	1
17	4.62×10^{-6}	.0131	1	1	1
18	0	.0046	.9996	1	1
19	0	.0015	.9812	1	1
20	0	.00045	.8957	1	1
21	0	.00013	.7323	1	1
22	0	.00003	.5365	.9949	1
23	0	7.66×10^{-6}	.3582	.9533	1
24	0	1.696×10^{-6}	.2218	.8433	1
25	0	0	.1290	.6782	1
26	0	0	.0710	.4990	1
27	0	0	.0372	.3405	.9970
28	0	0	.0186	.2182	.9701
29	0	0	.0089	.1327	.8917
30	0	0	.00405	.0770	.7602
31	0	0	.0018	.0428	.6009
32	0	0	.0007	.0228	.4439
33	0	0	.0003	.0117	.3096
34	0	0	.0001	.0058	.2055
35	0	0	.00004	.0028	.1308
36	0	0	.00001	.0013	.0801
37	0	0	5.05×10^{-6}	.00056	.0474
38	0	0	1.64×10^{-6}	.00024	.0271
39	0	0	0	.0001	.0150
40	0	0	0	.00004	.0080
41	0	0	0	.00001	.0042

$P(\max\{f_1, f_2, \cdots, f_K\} \geq x)$ $(K = 7)$

x	$n = 140$	$n = 175$	$n = 200$	$n = 225$	$n = 250$
27	.4027	.9991	1	1	1
28	.2650	.9855	1	1	1
29	.1649	.9321	1	1	1
30	.0977	.8238	1	1	1
31	.0555	.6746	.9974	1	1
32	.0303	.5142	.9773	1	1
33	.0159	.3683	.9169	1	1
34	.0081	.2502	.8083	.9997	1
35	.0039	.1625	.6664	.9943	1
36	.0019	.1014	.5159	.9672	1
37	.0008	.0611	.3780	.9006	1
38	.0004	.0356	.2642	.7918	.9991
39	.00016	.0201	.1773	.6558	.9897
40	.000065	.0110	.1147	.5137	.9555
41	.00003	.0059	.0719	.3831	.8834
42	.00001	.0030	.0437	.2737	.7746
43	3.73×10^{-6}	.0015	.0258	.1885	.6437
44	1.35×10^{-6}	.0007	.0148	.1256	.5088
45	0	.0003	.0083	.0812	.3847
46	0	.00016	.0045	.0511	.2798
47	0	.00007	.0024	.0313	.1967
48	0	.00003	.0012	.0187	.1342
49	0	.00001	.0006	.0109	.0890
50	0	5.47×10^{-6}	.0003	.0062	.0576
51	0	2.20×10^{-6}	.00015	.0035	.0364
52	0	0	.00006	.0019	.0224
53	0	0	.00003	.001	.0135
54	0	0	.00001	.0005	.0080
55	0	0	6.09×10^{-6}	.00026	.0046
56	0	0	2.58×10^{-6}	.0001	.0026
57	0	0	1.07×10^{-6}	.00006	.0014
58	0	0	0	.00003	.0008
59	0	0	0	.00001	.0004
60	0	0	0	6.21×10^{-6}	.0002

$P(\max\{f_1, f_2, \cdots, f_K\} \geq x)$ $(K = 10)$

x	$n = 125$	$n = 150$	$n = 200$	$n = 250$	$n = 300$
18	.5958	.9903	1	1	1
19	.3864	.9301	1	1	1
20	.2264	.7859	1	1	1
21	.1225	.5882	1	1	1
22	.0622	.3952	.9999	1	1
23	.0298	.2433	.9963	1	1
24	.0136	.1397	.9676	1	1
25	.0059	.0757	.8812	1	1
26	.0024	.0390	.7336	1	1
27	.0010	.0192	.5573	1	1
28	.0004	.0091	.3904	.9983	1
29	.0001	.0041	.2558	.9833	1
30	.00005	.0018	.1585	.9302	1
31	.000015	.0007	.0937	.8241	1
32	4.96×10^{-6}	.0003	.0531	.6769	1
33	1.53×10^{-6}	.0001	.0291	.5173	.9992
34	0	.00004	.0153	.3710	.9907
35	0	.000016	.0078	.2522	.9568
36	0	5.53×10^{-6}	.0039	.1638	.8807
37	0	1.86×10^{-6}	.0019	.1023	.7624
38	0	0	.0009	.0617	.6193
39	0	0	.0004	.0361	.4745
40	0	0	.0002	.0205	.3453
41	0	0	.00007	.0113	.2405
42	0	0	.00003	.0061	.1613
43	0	0	.00001	.0032	.1046
44	0	0	4.65×10^{-6}	.0016	.0659
45	0	0	1.76×10^{-6}	.0008	.0403
46	0	0	0	.0004	.0241
47	0	0	0	.0002	.0140
48	0	0	0	.00008	.0080
49	0	0	0	.00004	.0044
50	0	0	0	.00002	.0024
51	0	0	0	7.12×10^{-6}	.0013

5 The Sparse Case

The sparse case corresponds to large K and comparatively smaller, and even much smaller, values of n. Exact P-values are reported in some selected sparse cases. Inspection of the P-values reveals an interesting phenomenon; the P-values drop suddenly. That is, numerous empty cells and significant clustering will typically manifest in sparse multinomial data, and a lot of caution is needed before declaring any deviation from uniformity.

$P(\max\{f_1, f_2, \cdots, f_K\} \geq x)$ $(K = 50)$

x	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
2	.6183	.9880	1	1	
3	.0429	.3153	.7169	.9468	.9965
4	.0015	.0298	.1385	.3556	.6296
5	.00004	.0019	.0150	.0578	.1522
6	0	.0001	.0013	.0068	.0238
7	0	3.95×10^{-6}	.00009	.0030	
8	0	0	5.05×10^{-6}	.00006	.0003
9	0	0	0	3.99×10^{-6}	.00003
10	0	0	0	0	2.53×10^{-6}

$K = 100$

x	$n = 15$	$n = 30$	$n = 50$	$n = 80$	$n = 100$
2	.6687	.9922	1	1	1
3	.0411	.2931	.7880	.9976	1
4	.0012	.0221	.1504	.6050	.8738
5	.00003	.0012	.0145	.1228	.2984
6	0	.00005	.0011	.0159	.0524
7	0	1.66×10^{-6}	.00007	.0017	.0071
8	0	0	3.69×10^{-6}	.00015	.0008
9	0	0	0	.00001	.00008
10	0	0	0	0	7.63×10^{-6}

$P(\max\{f_1, f_2, \cdots, f_K\} \geq x)$ ($K = 250$)

x	$n = 15$	$n = 30$	$n = 50$	$n = 80$	$n = 100$
2	.3484	.8368	.9948	1	1
3	.0070	.0586	.2432	.6683	.8780
4	.00008	.0016	.0127	.0769	.1707
5	0	.00003	.00047	.0048	.0140
6	0	0	.00001	.0002	.0009
7	0	0	0	.00001	.00005
8	0	0	0	0	2.2×10^{-6}

$K = 400$

x	$n = 20$	$n = 30$	$n = 50$	$n = 80$	$n = 100$
2	.3830	.6722	.9591	.9998	1
3	.0069	.0239	.1071	.3657	.5826
4	.00007	.0004	.0033	.0210	.0495
5	0	5.28×10^{-6}	.00008	.0008	.0024
6	0	0	1.41×10^{-6}	.00002	.0001
7	0	0	0	0	3.19×10^{-6}

6 References

Barbour, A., Holst, L. and Janson, S. (1992). Poisson Approximation, Clarendon Press, New York.

Diaconis, P. and Graham, R. (1981). The analysis of sequential experiments with feedback to subjects, Ann. Statist., 9, 3-23.

Hall, P. and Titterington, D. (1987). On smoothing sparse multinomial data, Austr. and New Zealand Jour. Stat., 29, 19-37.

Koehler, K. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials, JASA, 75, 336-344.

Kolchin, V., Sevas'tyanov, B. and Chistyakov, V. (1978). Random Allocations, V. H. Winston & Sons, Washington.

Levin, B. (1983). On calculations involving the maximum cell frequency, Comm. Statist., Theory & Methods, 12, 1299-1327.

Rukhin, A. (2006). Gamma distribution order statistics, maximal multinomial frequency and randomization designs, JSPI, 136, 2213-2226.

Simonoff, J. (1983). A penalty function approach to smoothing large sparse contingency tables, Ann. Statist., 11, 208-218.

Zelterman, D. (1987). Goodness of fit tests for large sparse multinomial distributions, JASA, 82, 624-629.