SOME NEW TABLES OF THE LARGEST ROOT OF A MATRIX IN MULTIVARIATE ANALYSIS: A COMPUTER APPROACH FROM 2 TO 6

William W. Chen, Internal Revenue Service
Presented at the 2002 American Statistical Association

The distribution of the non-null characteristic roots of a matrix derived from sample observations taken from multivariate normal populations is of fundamental of importance in multivariate analysis. The Fisher-Girshick-Shu-Roy distribution (1939), which has interested statisticians for more than 6 decades, is revisited. Instead of using K.C.S. Pillai's method by neglecting higher order terms of the cumulative distribution function (CDF) of the largest root to approximate the percentage points, we simply keep the whole CDF and apply its natural nondecreasing property to calculate the exact probabilities. At the duplicated percentage points, we found our computed percentage points consistent with the existing tables. However, our tabulations have greatly extended the existing tables.

INTRODUCTION

We are concerned here with the distribution of the largest characteristic roots in multivariate analysis, when there are roots that range from 2 to 6. Fisher-Girshick-Shu-Roy (1939) discuss this in detail and present the exact joint probability density function in general. This well-known distribution depends on the number of characteristic roots and two parameters m and n. They are defined differently for various situations as described by Pillai (1955, 1957). The upper percentage points of the distribution are commonly used in three different types of hypothesis testing in multivariate analysis, namely: i) test of equality of the variance-covariance matrices of two p-variate normal populations; ii) test of equality of the p-dimensional mean vectors for k p-variate normal populations; and iii) test of independence between a p-set and a q-set of variates in a (p+q)-variate normal population. When the null hypotheses to be tested are true, all three types of test proposed above have been shown to depend only on the characteristic roots of matrices using observed samples. We could state the problem in the following manner. Using a random sample from the multivariate normal population, we could compute the characteristic roots from a usual sum of product matrices of this sample. We then compare the largest characteristic root of the matrices with the percentage points that we have tabulated in this paper to determine whether or not to reject the null hypothesis at a certain probability confidence. For this reason, the percentage points of the largest characteristic roots of the distribution have seriously attracted the attention of mathematical statisticians for more than 6 decades. There are already many published tables that either focus on upper percentage point tabulations or chart the various sizes of roots. K.C.S. Pillai is the most well known contributor in this area. He gave the general rules for finding the CDF of the largest root and tabulated upper percentage points of 95 percent and 99 percent for various root sizes. Other contributors, including D.N. Nanda (1948, 1951), F.G. Foster (1957, 1958), D.H. Rees (1957), and D.L. Heck (1960), will be discussed in more detail later. We will also discuss in detail the algorithm used to create tables for this paper. We will then compare the K.C.S. Pillai method with ours and also the advantage in our approach. The appendix lists the CDF's from 2 to 6.

CUMULATIVE FUNCTION AND HISTORICAL WORK

The joint distribution of s non-null characteristic roots of a matrix in multivariate distribution was given by Fisher-Girshick-Hsu-Roy (1939) (see the list of CDF's from 2 to 6 in the appendix). In this study, we were interested in the distribution of the largest characteristic root with the given CDF from 2 to 6. Even though we know the form of the joint density function, it may not be easy to write out the CDF of the largest characteristic root. There are two methods to find the CDF more easily. K.C.S. Pillai 1965) suggested that the CDF of the largest characteristic root could be presented in the determinantal form of incomplete beta functions. To overcome the difficulty of numerical integration of each of the s! multiple integrals when the determinant is expanded, he suggested an alternative reduction formula. This formula gives an exact expression for the CDF of the largest root in terms of incomplete beta functions or functions of incomplete beta functions for various values of s. Later, Pillai (1956b) expanded the CDF by neglecting higher order terms and tabulated the 95-percent and 99-percent percentage points. An alternative method suggested by D.N. Nanda (1948) yielded the same results. He started with the Vandermonde determinant and expanded it in minors of a row, then repeated applied integration by part to find the CDF of the largest characteristic root. In this paper, we slightly modified the D.N. Nanda notation and presented the case with roots ranging from 2 to 6. Following these CDF's and the algorithm described later, we could tabulate the upper percentage points.

It is useful here to review some of the published tables and see some reasons to extend the tables. K.C.S. Pillai (1956a, 1957, 1959) published tables that focus only on two percentage points, i.e., 95 percent and 99 percent for s =2,6, m = 0(1)4, and n varying from 5 to 1000. Foster and Rees (1957) tabulated the upper percentage points 80 percent, 85 percent, 90 percent, 95 percent, and 99 percent of the largest root for s=2, m=-0.5, 0(1)9, n=1(1)19(5)49,59, 79. Foster (1957, 1958) further extended these tables for values of s=3 and 4. Heck (1960) has given some charts of upper 95-percent, 97.5-percent, and 99-percent points for s=2(1)5, m=-0.5, 0(1)10, and n greater than 4.

Without a modern computer, it used to be an understandably difficulty task to compute the whole CDF(3.2) at each percentage point. This is not only tedious but worthless. Therefore, deleting higher order terms and keeping a few lower order terms to approximate the roots will form a good and reasonable method for solving the problem. But this approach involves intolerable error at the lower percentage points, such as 80 percent, 82.5 percent, 85 percent, 87.5 percent, 90 percent, or 92.5 percent. These percentage points are usually ignored, not because of lack of use but because of the difficulty of computation. Traditional methods treat missing values by interpolation. However, without say 85-percent or 90-percent points, it is difficult to interpolate 87.5 percent. In recent years, the computer has gradually matured in memory, speed, and flexibility. It has greatly changed the method we use for analyzing statistics. In this study, we use one of the most basic properties of the CDF and revisit this most important distribution. We attempted to include as many percentage points as we needed in one computer run. The upper percentage points we included are 0.80, 0.825, 0.850, 0.875, 0.890, 0.900, 0.910(0.005), and 0.995. Different authors have selected different m and n parameter values. We selected these two parameters in such a way that all existed table values will be included. For the parameter m=0(1)15 and the parameter n=1(1)20(2)30(5)80(10)150,200(100)1000, our table will give us the exact accuracy percentage points and probabilities and avoid the interpolation problem.

THE ALGORITHM

In this section, we describe in more detail how we compute the percentage points. For this study, no new theory was created. Instead, we applied the fundamental nondecreasing function property of the CDF, i.e., $\textit{if}\ x_1 \le x_2, \text{then}\ f(x_1) \le f(x_2)$. Applying this useful and simple property helps us find all the needed percentage points. Let us start with a standard procedure used in computer algorithms to see how we generate one percentage point. First, choose one set of m and n values, say m = 1 and n =2, and a very small x value, say 0 or $0.1*10^{-4}$ to ensure that there are no missing percentage points we are interested in that are larger than this value. Using these selected values, substitute into the equation (3.2) to compute the probability cumulate to this selected x value. If the computed probability equals, say 0.95000325, then write out this computed probability, m, n, and x values in a specified file, say f950.dat. Then, loop the pointer back and add a very small amount on x, say $0.1*10^{-4}$, and again compute the probability. If this time the computed probability is 0.9600125, then write out this computed probability, m, n, and x values in a different specified file, say f960.dat. Since we know that the cumulative function is always nondecreasing and continuous, it ensures us that any probability ranged from 0 to 1 will have a chance to be reached at least once for some selected x values. It is possible for several specific x values to round to the same probability. This means that we could increase either m or n by a selected value and reset x to 0 or a small value again to repeat the process of adding a small amount to x to compute the corresponding probability. This process should continue until we fill all m by n tables. Our experience shows that, for a chosen fixed m and n, as x increases by the above-stated increment, the computed probabilities also increase with multiple values rounding to the desired probability. The following simple rule has been adopted to select a triplet x, m, and n, for a desired probability. Let us say the desired probability y is p_0 and the estimate for x to reach this probability y is x_0: $\Pr(\theta_5 \le x_0) = p_0$. We need to find a pair say, x_0' and x_0'' such that :

$$\Pr(\theta_5 \le x_0') < p_0 < \Pr(\theta_5 \le x_0'')$$

We then can conclude by monotonicity that x_0, in the interval (x_0', x_0''), is the desired estimated ordinate x and report in the attached table. In the attached table, we have rounded our results to four decimal accurate places.

SOME CONCLUDING REMARKS

Pillai's approximation method by neglecting higher order terms has some limitations. Pillai (1954) studied these limitations in more detail for the case s =2,3, and 4. If we define the error of approximation of the upper percentage points of the distribution as the difference between the approximate and exact probabilities, then his comparative study obtained the following conclusions: i) There is greater agreement between the probabilities for the approximate and exact cases in the upper 99-percent points than in the 95-percent; ii) The difference between the approximate and exact probabilities in the upper 95-percent points occurs in the fifth decimal place; that on rounding gives a difference of only one in the fourth decimal place; iii) If we fixed the parameter m as constant, the error of approximation increases slowly as the other parameter increased; such increase occurs only in the sixth decimal place or at most is unity in the fifth decimal place when rounding.

Pillai (1959) also concluded that the approximate formula is only appropriate for percentage points 95 percent or higher. It might be adequate for those percentage points slightly below 95 percent. In application, it is very clear that lower percentage points are needed. Using the algorithm suggested in section 4, we can compute any percentage points. Since our method used the whole distribution function and not a truncated distribution, the table included in this paper is only a small portion of the table generated by computer. Interested readers may write to the author for more detailed tabulations.

ACKNOWLEDGMENT

The author is grateful to William Wong and Paul McMahon for their careful proofreading before this manuscript was submitted for publication.

REFERENCES

[1] Fisher, R.A. (1939), The sampling distribution of some statistics obtained from non-linear equation, Ann. Eugenics, Volume 9, pp. 238-249.

[2] Foster, F.G. and Rees, D.H. (1957), Upper percentage points of the generalized beta distribution, I, Biometrika, Volume 44, pp. 237-247.

[3] Foster, F.G.(1957), Upper percentage points of the generalized beta distribution, II, Biometrika, Volume 44, pp. 441-453.

[4] Foster, F.G. (1958), Upper percentage points of the generalized beta distribution, III, Biometrika, Volume 45, pp. 492-503.

[5] Girshick, M.A. (1939), On the sampling theory of the roots of determinantal equations, Ann. Math. Stat., Volume 10, pp. 203-224.

[6] Heck, D.L. (1960), Charts on some upper percentage points of the distribution of the largest characteristic root, Ann. Math. Stat., 31, pp. 625-642.

[7] Hsu, P.L. (1939), On the distribution of roots of certain determinantal equations, Ann. Eugenics, Volume 9, pp. 250-258.

[8] Nanda, D.N. (1948), Distribution of a root of a determinantal equation, Ann. Math. Stat., Volume 19, pp. 47-57.

[9] Nanda, D.N. (1951), Probability distribution tables of the largest root of a determinantal equation with two roots, J. Indian Soc. Of Agricultural Stat., Volume 3, pp. 175-177.

[10] Pillai, K.C.S. (1955), Some new test criteria in multivariate analysis, Ann. Math. Stat., 26, pp. 117-121.

[11] Pillai, K.C.S. (1956a), On the distribution of the largest or smallest root of a matrix in multivariate analysis, Biometrika, Volume 43, pp. 122-127.

[12] Pillai, K.C.S. (1956b), Some results useful in multivariate analysis, Ann. Math. Stat., Volume 27, pp. 1106-1114.

[13] Pillai, K.C.S. (1957), Concise tables for Statisticians, The Statistical Center University of the Philippines, Manila.

[14] Pillai, K.C.S. and Bantegui, C.G. (1959), On the distribution of the largest of Six roots of a matrix in multivariate analysis, Biometrika, Volume 46, pp. 237-240.

[15] Roy, S.N.(1939), p-statistics or some generalizations in analysis of variance appropriate to multivariate problems, Sankhya, Volume 4, pp. 381-396.

[16] Roy, S.N. (1945), The individual sampling distribution of the maximum, the minimum, and any intermediate of the p-statistics on the null hypothesis, Sankhya, Volume 7, pp. 133-158.

Upper percentage points of .900 of theta(p,m,n),
the largest eigenvalue of |B-theta(W+B)|=0,when s=2

	m							
n	0	1	2	3	4	5	6	7
1	.8464	.8968	.9221	.9374	.9476	.9550	.9605	.9649
2	.7307	.8058	.8474	.8742	.8928	.9067	.9173	.9258
3	.6366	.7244	.7768	.8120	.8375	.8568	.8719	.8842
4	.5618	.6551	.7138	.7548	.7853	.8090	.8278	.8433
5	.5017	.5965	.6587	.7035	.7375	.7644	.7862	.8042
6	.4527	.5468	.6106	.6577	.6942	.7234	.7475	.7676
7	.4122	.5043	.5685	.6169	.6550	.6860	.7117	.7334
8	.3782	.4677	.5315	.5805	.6196	.6517	.6787	.7016
9	.3493	.4359	.4989	.5479	.5875	.6204	.6483	.6721
10	.3244	.4080	.4698	.5186	.5584	.5918	.6202	.6448
11	.3028	.3834	.4439	.4921	.5319	.5655	.5943	.6194
12	.2839	.3616	.4206	.4681	.5077	.5413	.5704	.5958
13	.2671	.3421	.3996	.4463	.4855	.5191	.5482	.5739
14	.2523	.3245	.3805	.4264	.4651	.4985	.5276	.5534
15	.2390	.3087	.3632	.4082	.4464	.4794	.5085	.5342
16	.2270	.2943	.3473	.3914	.4290	.4617	.4906	.5163
17	.2161	.2812	.3328	.3759	.4129	.4453	.4739	.4995
18	.2063	.2692	.3194	.3616	.3980	.4299	.4583	.4837
19	.1973	.2581	.3070	.3483	.3840	.4155	.4436	.4689
20	.1890	.2480	.2956	.3359	.3711	.4021	.4299	.4549
22	.1744	.2299	.2750	.3137	.3475	.3776	.4047	.4293
24	.1619	.2142	.2571	.2941	.3267	.3559	.3823	.4063
26	.1510	.2005	.2414	.2769	.3083	.3365	.3622	.3857
28	.1416	.1885	.2275	.2615	.2918	.3191	.3441	.3670
30	.1332	.1778	.2151	.2478	.2770	.3034	.3277	.3500
35	.1161	.1557	.1893	.2189	.2457	.2702	.2927	.3137
40	.1028	.1385	.1690	.1961	.2208	.2434	.2645	.2841
45	.0923	.1248	.1526	.1776	.2004	.2215	.2412	.2596
50	.0837	.1135	.1391	.1623	.1835	.2032	.2216	.2390

Upper percentage points of .900 of theta(p,m,n),
the largest eigenvalue of |B-theta(W+B)|=0,when s=4

	m							
n	0	1	2	3	4	5	6	7
1	.9394	.9545	.9636	.9696	.9739	.9772	.9797	.9817
2	.8744	.9022	.9198	.9319	.9409	.9477	.9531	.9575
3	.8095	.8473	.8723	.8901	.9036	.9140	.9224	.9293
4	.7497	.7947	.8255	.8481	.8654	.8791	.8902	.8995
5	.6961	.7460	.7812	.8075	.8280	.8445	.8580	.8694
6	.6485	.7016	.7399	.7691	.7923	.8110	.8267	.8398
7	.6063	.6614	.7019	.7333	.7585	.7792	.7965	.8113
8	.5688	.6250	.6670	.7000	.7268	.7490	.7678	.7838
9	.5354	.5920	.6350	.6692	.6972	.7206	.7405	.7577
10	.5055	.5621	.6056	.6406	.6695	.6939	.7148	.7329
11	.4786	.5348	.5786	.6142	.6437	.6688	.6904	.7093
12	.4543	.5100	.5538	.5896	.6196	.6453	.6675	.6870
13	.4323	.4873	.5309	.5668	.5971	.6232	.6459	.6658
14	.4123	.4664	.5097	.5456	.5761	.6024	.6254	.6458
15	.3940	.4472	.4901	.5258	.5564	.5829	.6062	.6268
16	.3773	.4295	.4718	.5074	.5379	.5645	.5880	.6088
17	.3619	.4131	.4549	.4901	.5206	.5472	.5707	.5918
18	.3476	.3978	.4390	.4740	.5042	.5308	.5544	.5756
19	.3345	.3837	.4243	.4588	.4889	.5154	.5390	.5602
20	.3222	.3704	.4104	.4446	.4744	.5008	.5243	.5456
22	.3003	.3465	.3852	.4185	.4478	.4738	.4972	.5184
24	.2811	.3255	.3629	.3953	.4239	.4495	.4727	.4937
26	.2642	.3068	.3430	.3745	.4024	.4276	.4504	.4712
28	.2492	.2902	.3251	.3557	.3830	.4076	.4300	.4506
30	.2358	.2752	.3090	.3387	.3653	.3894	.4114	.4316
35	.2079	.2438	.2748	.3024	.3274	.3501	.3711	.3905
40	.1858	.2187	.2474	.2732	.2965	.3180	.3379	.3564
45	.1680	.1983	.2250	.2490	.2710	.2912	.3101	.3277
50	.1533	.1814	.2063	.2288	.2494	.2686	.2865	.3033

Upper percentage points of .900 of theta(p,m,n),
the largest eigenvalue of |B-theta(W+B)|=0,when s=5

	m							
n	0	1	2	3	4	5	6	7
1	.9568	.9664	.9725	.9767	.9798	.9822	.9840	.9856
2	.9062	.9249	.9373	.9462	.9529	.9580	.9622	.9656
3	.8526	.8793	.8976	.9111	.9213	.9295	.9361	.9415
4	.8008	.8338	.8571	.8746	.8882	.8991	.9080	.9155
5	.7526	.7903	.8177	.8385	.8550	.8684	.8794	.8888
6	.7085	.7497	.7802	.8038	.8226	.8381	.8511	.8621
7	.6684	.7121	.7449	.7707	.7916	.8089	.8234	.8359
8	.6321	.6774	.7120	.7395	.7620	.7808	.7968	.8105
9	.5991	.6455	.6814	.7102	.7340	.7541	.7712	.7860
10	.5691	.6161	.6529	.6828	.7076	.7287	.7468	.7626
11	.5418	.5891	.6265	.6571	.6827	.7046	.7236	.7401
12	.5168	.5641	.6019	.6330	.6593	.6819	.7015	.7187
13	.4940	.5411	.5790	.6105	.6373	.6603	.6805	.6983
14	.4730	.5198	.5577	.5894	.6165	.6400	.6605	.6788
15	.4536	.5000	.5378	.5696	.5969	.6207	.6416	.6602
16	.4358	.4816	.5192	.5510	.5785	.6024	.6236	.6425
17	.4192	.4644	.5018	.5336	.5610	.5852	.6066	.6257
18	.4038	.4484	.4855	.5171	.5446	.5688	.5903	.6096
19	.3895	.4335	.4701	.5016	.5290	.5532	.5748	.5943
20	.3762	.4194	.4557	.4869	.5142	.5384	.5601	.5797
22	.3520	.3939	.4293	.4600	.4870	.5110	.5327	.5524
24	.3307	.3712	.4057	.4357	.4623	.4862	.5078	.5274
26	.3118	.3510	.3845	.4139	.4400	.4636	.4849	.5045
28	.2950	.3328	.3654	.3941	.4197	.4429	.4640	.4834
30	.2798	.3164	.3480	.3760	.4012	.4239	.4448	.4639
35	.2479	.2816	.3111	.3373	.3611	.3829	.4029	.4214
40	.2225	.2537	.2811	.3058	.3283	.3489	.3680	.3858
45	.2018	.2308	.2564	.2796	.3008	.3204	.3387	.3558
50	.1847	.2116	.2357	.2575	.2776	.2962	.3136	.3300

Upper percentage points of .900 of theta(p,m,n),
the largest eigenvalue of |B-theta(W+B)|=0,when s=6

	m							
n	0	1	2	3	4	5	6	7
1	.9677	.9741	.9784	.9815	.9838	.9856	.9870	.9882
2	.9272	.9404	.9496	.9562	.9614	.9654	.9687	.9714
3	.8824	.9020	.9159	.9263	.9344	.9408	.9461	.9506
4	.8376	.8625	.8806	.8944	.9053	.9142	.9215	.9276
5	.7946	.8238	.8455	.8623	.8757	.8867	.8959	.9037
6	.7542	.7868	.8114	.8307	.8464	.8593	.8702	.8795
7	.7168	.7519	.7788	.8003	.8178	.8324	.8448	.8555
8	.6822	.7192	.7480	.7712	.7903	.8063	.8201	.8319
9	.6503	.6888	.7190	.7435	.7640	.7812	.7961	.8090
10	.6210	.6604	.6917	.7174	.7389	.7572	.7730	.7869
11	.5939	.6340	.6661	.6927	.7151	.7342	.7509	.7655
12	.5690	.6094	.6421	.6693	.6924	.7124	.7297	.7450
13	.5459	.5865	.6196	.6474	.6710	.6915	.7095	.7254
14	.5245	.5651	.5985	.6266	.6507	.6717	.6901	.7065
15	.5046	.5452	.5787	.6070	.6315	.6528	.6717	.6885
16	.4862	.5265	.5600	.5885	.6132	.6349	.6541	.6712
17	.4689	.5090	.5424	.5710	.5959	.6178	.6372	.6547
18	.4529	.4926	.5259	.5545	.5794	.6015	.6212	.6389
19	.4378	.4771	.5103	.5388	.5638	.5860	.6058	.6237
20	.4237	.4626	.4955	.5240	.5490	.5712	.5912	.6092
22	.3980	.4359	.4683	.4965	.5214	.5437	.5637	.5820
24	.3752	.4121	.4438	.4716	.4963	.5185	.5386	.5569
26	.3548	.3907	.4218	.4491	.4735	.4955	.5155	.5338
28	.3365	.3714	.4017	.4285	.4526	.4744	.4942	.5124
30	.3199	.3539	.3835	.4097	.4334	.4549	.4746	.4927
35	.2849	.3164	.3442	.3691	.3917	.4124	.4315	.4491
40	.2567	.2861	.3122	.3358	.3573	.3771	.3954	.4125
45	.2335	.2610	.2856	.3078	.3283	.3472	.3648	.3813
50	.2142	.2400	.2631	.2842	.3036	.3217	.3386	.3544