AD-A065 738 ROCHESTER UNIV NY DEPT OF STATISTICS F/G 12/1
A GAUSSIAN APPROXIMATION TO THE DISTRIBUTION OF SAMPLE VARIANCE--ETC(U)
1979 G S MUDHOLKAR, M C TRIVEDI AFOSR-77-3360
UNCLASSIFIED AFOSR-TR-79-0251 NL

| OF |
AD
A065738

END
DATE
FILMED
5--79
DDC

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD A0 65738

DDC FILE COPY

LEVEL

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR-79-0251
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): A GAUSSIAN APPROXIMATION TO THE DISTRIBUTION OF SAMPLE VARIANCE FOR NONNORMAL POPULATIONS
5. TYPE OF REPORT & PERIOD COVERED: Interim rept.
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Govind S. Mudholkar ~~and~~ Mudhusudan C. Trivedi
8. CONTRACT OR GRANT NUMBER(s): AFOSR-77-3360
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Rochester, Department of Statistics, Rochester, New York 14627
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332
12. REPORT DATE: 1979
13. NUMBER OF PAGES: 15
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 16p.
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Gaussian Approximation, Sample Variance, Nonnormal Parent Populations
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A Gaussian approximation to the distribution of sample variance using Wilson-Hilferty approach is developed. It is studied for accuracy and compared with the well known approximations due to Box and Roy and Tiku by taking the exponential, the double exponential, the uniform, the product normal distributions as the parent populations. The Wilson-Hilferty approximation which can be used for probabilities as well as percentiles is seen to compare favorably with the other two approximations.

DD FORM 1 JAN 73 1473

410 276

UNCLASSIFIED

AFOSR-TR- 79-0251

A GAUSSIAN APPROXIMATION TO THE DISTRIBUTION OF SAMPLE VARIANCE FOR NONNORMAL POPULATIONS

by

Govind S. Mudholkar* and Madhusudan C. Trivedi
University of Rochester — Pennwalt Corporation

ABSTRACT

A Gaussian approximation to the distribution of sample variance using Wilson-Hilferty [12] approach is developed. It is studied for accuracy and compared with the well known approximations due to Box [2] and Roy and Tiku [8] by taking the exponential, the double exponential, the uniform, the product normal and various mixtures of normal distributions as the parent populations. The Wilson-Hilferty approximation which can be used for probabilities as well as percentiles is seen to compare favorably with the other two approximations.

Key Words: Gaussian Approximation, Sample Variance, Nonnormal Parent Populations.

*Research supported in part by the Air Force Office of the Scientific Research, Air Force Systems Command, USAF under Grant No. AFOSR-77-3360. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation hereon.

1. INTRODUCTION AND SUMMARY

Let $X_1, X_2, \ldots, X_n$ be a sample from F. Let $\overline{X} = \Sigma X_i/n$ and $S^2 = \Sigma(X_i-\overline{X})^2$. S^2 is a very commonly encountered statistic but its exact distribution is generally intractable except in a few cases such as a normal parent population or a mixture of normal populations. If F is a mixture of two normal populations differing only in means then Hyrenious [3] gives the exact distribution of S^2 as a binomial mixture of noncentral chisquare distributions. On the other hand if F is a mixture of two normal distributions with common mean but different variances then S^2 can be shown (see Appendix) to be distributed according to a binomial mixture of quadratic form distributions. The distribution of S^2 is otherwise unavailable but a number of approximations for it are known. The prominent among these are the scaled chisquare approximation due to Box [2] and the Laguerre polynomial series approximation by Roy and Tiku [8], which are as follows:

The Box Approximation. Box, in 1953, suggested approximating the distribution of $Y = S^2/C_2$, $C_2 = \mathrm{Var}(X)$, by a scaled chisquare variate in which the parameters are obtained by using the first two moments. Specifically,

$$\Pr(Y \le t) \approx \frac{1}{\Gamma(b)\rho^b} \int_0^t y^{b-1} e^{-y/\rho} \, dy \,, \tag{1.1}$$

where $\rho = \mathrm{Var}(Y)/m$, $b = m/\rho$, and $m = E(Y) = n-1$.

The Roy and Tiku Approximation. Roy and Tiku, in 1962, suggested use of Laguerre polynomials to derive a series approximation for the distribution of $Y = S^2/(2C_2)$. They proposed,

$$\Pr(Y \leq t) \approx \int_o^t P_m(y) \sum_{j=o}^{k} a_j^{(m)} L_j^{(m)}(y)\, dy, \tag{1.2}$$

where $P_m(y) = \frac{1}{\Gamma(m)} y^{m-1} e^{-y}, \; y \geq o,$

$$L_j^{(m)}(y) = \frac{1}{j!} \sum_{i=o}^{j} \binom{j}{i} (-y)^i \, \Gamma(m+j)/\Gamma(m+i), \tag{1.3}$$

is a Laguerre polynomial of degree j, $j \geq o$, $m = E(Y)$, k = number of terms in the approximation, and a_j are constants determined by using the first j moments. Actually,

$$a_j^{(m)} = \Gamma(m) \sum_{i=o}^{j} \binom{j}{i} E(-Y)^i/\Gamma(m+i). \tag{1.4}$$

Tan and Wong [11] show that the Roy and Tiku approximation can yield very unreasonable results in case of a very nonnormal parent population such as the exponential, the double exponential, or the product normal distribution. They also examine the two approximations and an alternative series approximation introduced by them in some detail when F is a mixture of two normal distributions with a common variance and different means. They find that the Roy and Tiku approximation and their alternative series approximation are superior to the Box approximation. It may be noted that neither the Roy-Tiku nor the Tan-Wong series approximations are very convenient for approximating percentiles.

In this paper the approach of E. Wilson and M. Hilferty [12] to approximating a chisquare distribution, which was later extended by Sankaran [9] and by Jensen and Solomon [5] to other cases, is adapted for developing a Gaussian approximation for S^2. The new approximation is presented in section 2. In section 3, this approximation is compared with the approximations due to Box [2] and Roy and Tiku [8] over a spectrum of parent populations, namely, various mixtures of normal distributions, the exponential, the double exponential, the uniform, and the product normal populations. The conclusions of the numerical study are summarized in section 4. The Wilson-Hilferty approximation is found to yield a reasonably good and generally superior approximation.

2. THE WILSON-HILFERTY APPROXIMATION

Given a nonnegative random variable Y the Wilson-Hilferty approach consists in obtaining an almost symmetrically distributed power Y^h of Y and approximating it by a Gaussian random variable. This reasoning may be attributed to Sankaran [9] who taking a cue from the Wilson-Hilferty approximation for a chisquare distribution developed an approximation for the noncentral chisquare distribution. It was further abstracted and extended to central and noncentral quadratic form distributions by Jensen and Solomon [5]. It may be summarized as follows.

Let κ_1, κ_2, ... denote the cumulants of Y and let $\phi_r = \kappa_r/\kappa_1$, r = 2,3,.. be bounded. Then by using the Taylor expansion we get,

$$\mu_1(h) = 1 + \frac{h(h-1)\phi_2}{2\kappa_1} + \frac{h(h-1)(h-2)}{24\kappa_1^2}[4\phi_3 + 3(h-3)\phi_2^2] + O(\kappa_1^{-3}) \qquad (2.1)$$

From this the r^{th} moment $\mu_r'(h) = E\,[(Y/\kappa_1)^h]^r$ is obtained by substituting rh for h. Simple computations then yield the following series expressions for these moments in terms of the powers of $(\kappa_1)^{-1}$ as follows.

$$\mu_2(h) = \frac{h^2\phi_2}{\kappa_1} + \frac{h^2(h-1)}{2\kappa_1{}^2}\,[2\phi_3 + (3h-5)\phi_2^2\,] + O(\kappa_1{}^{-3}), \tag{2.2}$$

$$\mu_3(h) = \frac{h^3}{\kappa_1{}^2}\,[\phi_3 + 3(h-1)\phi_2^2\,] + O(\kappa_1{}^{-3}), \tag{2.3}$$

$$\mu_4(h) = 3h^4\phi_2^2\,/\kappa_1{}^2 + O(\kappa_1{}^{-3}). \tag{2.4}$$

The exponent $h = h_o$ which approximately symmetrizes Y obtained by equating the leading term of $\mu_3(h)$ to zero is, therefore,

$$h_o = 1 - \kappa_1\,\kappa_3/3\kappa_2{}^2. \tag{2.5}$$

$(Y/\kappa_1)^{h_o}$ may now be approximated by the normal distribution with mean $\mu(h_o)$ and variance $\sigma^2(h_o) = \mu_2(h_o)$ given by (2.1) and (2.2) respectively.

Now let $X_1, X_2, \ldots, X_n$ be the random sample of size n from a population F with finite cumulants $C_1, C_2, \ldots$. Then it is well known (Kendall and Stuart page 290 [6]) that the cumulants κ_r, r = 1,2,3 of $Y = S^2/\sigma^2$ ($\sigma^2 = C_2$) are,

$$\begin{aligned}
\kappa_1 &= (n-1)\\
\kappa_2 &= (n-1)^2\,[C_4/(\,n\sigma^4\,) + 2/(n-1)]\\
\kappa_3 &= (n-1)^3\,[C_6/n^2 + 12C_4C_2/\{n(n-1)\} + 4(n-2)C_3^2/\{n(n-1)\}\\
&\quad + 8C_2^3/(n-1)^2\,]/\sigma^6.
\end{aligned} \tag{2.6}$$

It is easy to see that in this case $\phi_r = \kappa_r/\kappa_1$ are bounded and the Wilson-Hilferty approach is applicable. The exponent h_o is then obtained by (2.5) and $\mu(h_o)$ and $\sigma^2(h_o) = \mu_2(h_o)$ as described in (2.1) and (2.2) respectively. The resulting approximation to the distribution function of S^2 is then given by,

$$\Pr(S^2 \leq t) \approx \Phi[\{ (t/\kappa_1)^{h_o} - \mu(h_o)\}/\sigma(h_o)]. \tag{2.7}$$

The corresponding approximation to the α^{th} percentile of S^2 is,

$$S_\alpha^2 \approx \kappa_1[Z_\alpha\sigma(h_o) + \mu(h_o)]^{1/h_o} \tag{2.8}$$

where Z_α is the α^{th} percentile of standard normal distribution.

3. NUMERICAL COMPARISONS

This section contains numerical comparisons of the Wilson-Hilferty approximation for the distribution of S^2 with the scaled chisquare approximation due to Box [2] and the Laguerre polynomial series approximation due to Roy and Tiku [8]. The comparisons are made by either computing or simulating the true distributions of S^2 of samples from various nonnormal populations as described below.

3a. Mixture of Normal Distributions

Case 1. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with p.d.f.

$$f(x) = pN(\mu_1,\sigma^2) + (1-p)N(\mu_2,\sigma^2), \tag{3.1}$$

where $0 \leq p \leq 1$, $\sigma^2 > 0$, $-\infty < \mu_1, \mu_2 < \infty$ and $N(\mu,\sigma^2)$ denotes the normal density function with mean μ and variance σ^2. Then Hyrenius [3] has shown that,

$$\Pr(S^2/\sigma^2 \leq t) = \sum_{i=0}^{n} \binom{n}{i} p^i(1-p)^{n-i} \Pr(\chi'^2_{n-1}(\lambda_i) \leq t), \tag{3.2}$$

where $\chi'^2_{n-1}(\lambda_i)$ denotes the noncentral chisquare variable with n-1 degrees of freedom and the noncentrality parameter $\lambda_i = i(n-i)(\mu_1 - \mu_2)^2/(n\sigma^2)$. A selection of the values of the exact c.d.f., computed using (3.2) and the IMSL subroutine MDCH, together with the errors of the three approximations computed according to (1.1), (1.2), and (2.7) appear in Table 1.

Case 2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with p.d.f.

$$f(x) = pN(\mu, \sigma_1^2) + (1-p)N(\mu, \sigma_2^2), \tag{3.3}$$

where $0 \leq p \leq 1$, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, $-\infty < \mu < \infty$, and $N(\mu, \sigma^2)$ denotes a normal density function with mean μ and variance σ^2. Then it is shown in Appendix that,

$$\Pr(S^2 \leq t) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \Pr(\textstyle\sum \lambda_j Y_j \leq t), \tag{3.4}$$

where as described in Appendix $\sum_j \lambda_j Y_j$ is a quadratic form in independently distributed normal variables . A selection of the values of the exact c.d.f. computed using (3.4) and the subroutine FQUAD [7] prepared from the technique derived by Imhof [4] and the errors of three approximations appear in Table 2.

3b. Other Nonnormal Populations

The other nonnormal populations used for the comparisons are (i) uniform, (ii) exponential, (iii) product normal, and (iv) double exponential. The exact distributions of the sample variances from these populations are not available. Therefore, the c.d.f.'s are estimated from the following Monte Carlo experiments.

Using the generator RANDU, supported by the Digital Equipment Corporation on PDP 11/70 computers, to generate U(0,1) random variables and transformations such as Box-Meuller [1] 5000 random samples of size 20 each from the four populations were obtained. From these samples the empirical c.d.f. of S^2 for each population was then constructed. This process was repeated seven times. For each selected value of S^2 the average of the seven values of the c.d.f. was used as the value of Monte Carlo c.d.f.. The following is a brief explanation of the method used to generate random samples for each population.

(i) Uniform (0,1): Use of RANDU subroutine.

(ii) Exponential (1): Obtain $U = U(0,1)$ then $X = -2\log(U)$.

(iii) Product normal: $X = Z_1Z_2$ where Z_i, $i = 1,2$ are i.i.d. N(0,1). Obtain U_1 and U_2 using RANDU then compute $X = -\log(U_1)\ \mathrm{Sin}(4\pi U_2)$.

(iv) Double exponential (0,1): Obtain $U = U(0,1)$ then $X = \log(2U)$ if $U < .5$, or $X = -\log[2(1-U)]$ otherwise.

A selection of the values of the empirical c.d.f. of S^2 of the samples from the four populations together with the errors of the three approximations appear in Table 3.

4. CONCLUSIONS

From the numerical studies described in the previous section the following conclusions are drawn. The abbreviations W-H, R-T, and Box connote the Wilson-Hilferty, the Roy and Tiku, and the Box approximations respectively.

1. From Table 1, corresponding to the mixture of two normal distributions differing in means only the following can be observed. (a) The three approximations are reasonable for small values of $|\mu_1 - \mu_2|$ but their quality deteriorates as the value of $|\mu_1 - \mu_2|$ increases. (b) As the value of p increases W-H improves and Box worsens. (c) W-H is substantially superior to Box and R-T when the value of $|\mu_1 - \mu_2|$ is large; when the value of $|\mu_1 - \mu_2|$ is small it is slightly inferior to R-T. Box is not better than W-H anywhere.

2. From Table 2, corresponding to the mixture of two normal distributions differing in variances only, the following can be observed. (a) All three approximations are reasonable over the range of parameters considered. (b) Box is superior to W-H and R-T when p is small and the ratio of variances is large. (c) R-T is superior to W-H and Box when p as well as the ratio of variances is small. (d) Otherwise W-H and Box are equally good.

3. The observations from Table 3 corresponding to the uniform, the exponential, the product normal, and the double exponential populations are as follows. (a) R-T is the poorest performing approximation, in general embarrassingly so. Clearly the improper estimates of the probabilities are due to truncation of the series after four terms. (b) W-H is the best of the three approximations. Its performance appears to be substantially superior in all four cases.

4. In summary, it is concluded that the Wilson-Hilferty approximation, derived in section 2, is a reasonable approximation over the spectrum of populations considered. In no case is W-H the the poorest of the three nor is it embarrassingly bettered by either of the other two approximations. When it is superior it is substantially so.

TABLE 1. Exact C.D.F. of S^2/σ^2 of Samples from $pN(\mu_1,\sigma^2) + (1-p)N(\mu_2,\sigma^2)$ and Errors* of the Approximations $\sigma^2 = 4$ and N = 11.

p = .1					p = .2					p = .3					p = .4				
t	(1)	(2)	(3)	(4)	t	(1)	(2)	(3)	(4)	t	(1)	(2)	(3)	(4)	t	(1)	(2)	(3)	(4)
$\mu_1 - \mu_2 = 4$																			
6	.0903	15	25	-3	6	.0441	16	-40	-7	9	.0914	-7	-66	-4	9	.0628	-2	-59	-7
8	.1999	-20	-4	1	8	.1086	-8	-53	-3	11	.1675	-16	-44	4	11	.1243	-8	-49	1
10	.3324	-42	-29	6	10	.1999	-29	-41	6	15	.3683	-13	34	15	15	.3056	-10	17	18
12	.4673	-39	-36	6	14	.4228	-25	23	12	17	.4762	-5	61	12	17	.4113	-6	47	17
14	.5904	-21	-27	3	16	.5339	-10	46	7	19	.5787	2	72	6	19	.5163	-1	63	12
16	.6947	0	-12	-1	18	.6348	4	56	1	21	.6708	8	67	-1	21	.6143	4	65	4
18	.7785	14	1	-4	20	.7217	14	52	-5	24	.7835	10	44	-8	24	.7391	7	48	-7
22	.8911	20	12	-3	23	.8233	17	34	-8	30	.9197	5	-3	-7	30	.8985	4	2	-11
26	.9505	10	9	0	29	.9385	7	-3	-3	33	.9542	1	-14	-3	33	.9409	2	-11	-7
32	.9865	-1	2	1	35	.9817	-1	-13	2	39	.9867	-2	-16	2	39	.9822	-1	-15	1
$\mu_1 - \mu_2 = 10$																			
5	.0342	367	-94	-246	17	.0911	-181	-559	-176	30	.0706	-26	-316	-25	38	.0485	7	-194	-57
11	.2035	-513	-845	-609	23	.1209	-89	-366	244	38	.1306	-11	-161	120	47	.1127	60	-62	262
17	.3090	-553	-581	524	29	.1903	-50	-108	467	53	.3232	55	420	321	53	.1845	78	104	439
29	.4721	104	472	457	41	.3582	5	403	246	59	.4284	51	526	267	65	.3991	22	337	398
35	.5759	158	544	-331	47	.4540	55	554	35	65	.5409	27	498	137	71	.5257	-26	316	193
41	.6740	149	471	-649	53	.5527	89	581	-133	71	.6510	-3	369	-25	77	.6481	-59	220	-47
47	.7577	129	351	-515	65	.7378	74	338	-256	77	.7498	-26	195	-170	86	.7550	-65	96	-245
59	.8797	68	100	28	77	.8746	11	11	-171	89	.8925	-28	-88	-283	95	.9017	-24	-81	-378
71	.9498	2	-64	166	83	.9200	-11	-96	-101	95	.9355	-16	-151	-249	101	.9431	-2	-110	-326
77	.9697	-17	-98	129	89	.9516	-21	-154	-34	107	.9804	3	-155	-101	107	.9689	11	-108	-235

*Error = (Approximate C.D.F. - Exact C.D.F.)x 10^4. (1) Exact C.D.F. $Pr(S^2/\sigma^2 \leq t)$, (3.2); (2) Error: Wilson-Hilferty Approximation (2.7); (3) Error: Box Approximation (1.1); (4) Error: Roy-Tiku Approximation (1.2).

TABLE 2. Exact C.D.F. of S'^2 of Samples from $pN(\mu_1,\sigma_1^2) + (1-p)N(\mu_2,\sigma_2^2)$ and Errors* of the Approximations.

$\mu_1 = \mu_2 = 0, \quad \sigma_1^2 = 1$

$p = .1, \sigma_2^2 = 2$					$p = .1, \sigma_2^2 = 8$					$p = .4, \sigma_2^2 = 2$					$p = .4, \sigma_2^2 = 8$				
t	(1)	(2)	(3)	(4)	t	(1)	(2)	(3)	(4)	t	(1)	(2)	(3)	(4)	t	(1)	(2)	(3)	(4)
										N = 11									
4	.0553	4	9	-1	4	.0643	4	2	-9	4	.0616	2	44	-1	4	.1117	-5	17	-42
5	.1129	-4	8	1	5	.1247	-6	0	1	5	.1227	-5	41	-1	5	.1833	-27	1	50
7	.2796	-13	0	1	7	.2910	-17	-3	17	7	.2925	-13	2	1	7	.3459	-39	-21	148
8	.3756	-12	-5	1	8	.3842	-15	-3	17	8	.3873	-12	-19	2	8	.4276	-33	-24	116
10	.5614	-3	-9	0	10	.5630	-4	-2	4	10	.5673	-4	-41	1	10	.5773	-12	-22	-14
13	.7746	7	-6	0	13	.7691	8	0	-11	13	.7714	6	-31	-2	13	.7499	12	-9	-96
16	.8975	6	-1	0	16	.8910	8	1	-7	16	.8906	7	-7	-1	16	.8610	16	0	-32
19	.9574	2	2	0	19	.9527	3	1	0	19	.9512	3	6	0	19	.9263	11	4	21
22	.9835	0	2	0	22	.9807	0	0	3	22	.9793	0	9	1	22	.9623	4	4	26
28	.9979	-1	1	0	28	.9972	-1	0	1	28	.9967	-1	4	0	28	.9909	-2	2	1
										N = 20									
11	.0798	-1	5	0	11	.0900	-2	-1	3	11	.0891	-1	27	0	8	.0448	8	11	-47
13	.1644	-5	3	0	13	.1783	-7	-1	10	13	.1789	-5	14	1	10	.1034	-5	4	26
15	.2823	-7	-2	1	15	.2926	-8	-1	11	15	.2950	-7	-7	2	13	.2358	-19	-8	130
17	.4137	-5	-5	0	17	.4197	-6	0	6	17	.4233	-6	-25	1	17	.4514	-13	-12	-3
19	.5446	-1	-7	0	19	.5455	-1	1	-3	19	.5492	-2	-33	0	20	.6045	-1	-8	-106
20	.6057	1	-7	0	20	.6044	1	1	-7	20	.6077	0	-33	-1	22	.6920	6	-4	-102
22	.7142	3	-5	-1	22	.7091	4	1	-9	22	.7122	3	-27	-1	25	.7971	10	0	-35
25	.8363	4	-2	0	25	.8289	5	1	-7	25	.8290	5	-12	-1	28	.8721	10	2	-25
28	.9134	3	1	-1	28	.9064	4	0	-1	28	.9051	4	0	0	33	.9453	4	3	34
33	.9740	0	2	0	33	.9700	0	0	1	33	.9684	1	8	1	40	.9855	-1	1	-2

*Error = (Approximate C.D.F. = Exact C.D.F.)x 10^4. (1) Exact C.D.F. $Pr(S^2 \leqq t)$, (3.4); (2) Error: Wilson-Hilferty Approximation (2.7); (3) Error: Box Approximation (1.1); (4) Error: Roy-Tiku Approximation (1.2).

TABLE 3. Monte Carlo C.D.F.* of S^2 of Samples of Size 20 from Various Populations and Errors** of the Approximations.

t	(1)	(2)	(3)	(4)	t	(1)	(2)	(3)	(4)
		Uniform					Exponential		
1.1	.0703	15	-86	101	6	.0496	-284	486	1503
1.2	.1252	10	-48	206	8	.1179	-397	500	5715
1.3	.2009	18	37	285	12	.3095	-314	169	5528
1.5	.4085	31	199	228	14	.4028	-146	32	-4578
1.6	.5282	13	197	82	18	.5715	81	-196	-18204
1.7	.6396	36	190	-35	21	.6719	165	-268	-9245
1.8	.7410	32	127	-154	27	.8134	164	-270	-11618
2.0	.8896	-1	-23	-253	34	.9039	100	-161	3049
2.1	.9335	0	-53	-209	42	.9523	61	-35	-2479
2.2	.9628	-5	-69	-144	50	.9763	25	12	-920
		Product-Normal					Double Exponential		
6	.0590	-250	392	1732	15	.0546	-110	240	439
8	.1308	-320	371	6733	19	.1234	-120	228	954
10	.2188	-283	270	10824	27	.3144	-8	61	-160
14	.4062	-91	-2	-4713	31	.4194	30	-56	-1473
16	.4929	-4	-112	-16790	35	.5192	45	-153	-2145
21	.6705	104	-254	-11665	39	.6092	39	-219	-1796
27	.8085	121	-221	12884	45	.7196	26	-242	-120
34	.9006	67	-128	3893	52	.8125	23	-187	1344
42	.9498	50	-10	-2599	63	.9043	-6	-92	807
50	.9755	15	20	-1015	76	.9568	-11	-5	-301

*Each C.D.F. is estimated on the basis of seven sets of 5000 samples.

** Error = (Approximate C.D.F. - Monte Carlo C.D.F.)x 10^4.

(1) Monte Carlo C.D.F. Pr($S^2 \leq t$), (see section 3b);

(2) Error: Wilson-Hilferty Approximation (2.7); (3) Error: Box Approximation (1.2); (4) Error: Roy-Tiku Approximation (1.2).

APPENDIX

THE DISTRIBUTION OF SAMPLE VARIANCE FOR A SCALED MIXTURE OF NORMAL POPULATIONS

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with probability density function (p.d.f.)

$$f(x) = pN(0,1) + (1-p)N(0,\sigma^2), \tag{A.1}$$

$0 \leq p \leq 1$ and $N(\mu,\sigma^2)$ denotes the normal density function with mean μ and variance σ^2. The corrected sum of squares may be expressed as a quadratic form in X's as,

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \underset{\sim}{X}'\underset{\sim}{A}\,\underset{\sim}{X} \tag{A.2}$$

where $\underset{\sim}{X}' = (X_1, X_2, \ldots, X_n)$, $\underset{\sim}{A} = (\underset{\sim}{I}_n - n^{-1} \underset{\sim}{J}_n)$, and $\underset{\sim}{J}_n$ is the n x n matrix of 1's. Using this representation it is easy to compute the characteristic function of $\underset{\sim}{X}'\underset{\sim}{A}\,\underset{\sim}{X}$ as given in the following proposition.

<u>Proposition</u>: The characteristic function of $\underset{\sim}{X}'\underset{\sim}{A}\,\underset{\sim}{X}$ is given by,

$$\Psi(t) = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} \left|\underset{\sim}{I} - 2it\, \underset{\sim}{A}\, \underset{\sim}{\Lambda}_r\right|^{-1/2}, \tag{A.3}$$

where $\underset{\sim}{\Lambda}_r$ is a matrix

$$\underset{\sim}{\Lambda}_r = \left(\begin{array}{c|c} \underset{\sim}{I}_r & \underset{\sim}{O} \\ \hline \underset{\sim}{O} & \sigma^2 \underset{\sim}{I}_{n-r} \end{array}\right). \tag{A.4}$$

The p.d.f. of S^2 can be obtained by inverting the above characteristic function. This may be done as follows, Let $\underset{\sim}{A}\,\underset{\sim}{\Lambda}_r = \underset{\sim}{B}_r = \underset{\sim}{B}$ which is a symmetric matrix of order n. Now suppressing the suffix r, there exists a nonsingular matrix $\underset{\sim}{T}$, such that, $\underset{\sim}{T}^{-1}\underset{\sim}{B}\,\underset{\sim}{T} = \text{diag}\,(\,\underset{\sim}{D}_1, \underset{\sim}{D}_2, \ldots, \underset{\sim}{D}_k\,) = \underset{\sim}{D}$, k = number of distinct eigenvalues λ_i of $\underset{\sim}{B}$ with respective multiplicity n_i, $\underset{\sim}{D}_i = \lambda_i \underset{\sim}{I}_{n_i}$, and $\sum n_i = n$. Thus,

$$|I - 2it\,\underset{\sim}{B}| = |\underset{\sim}{T}^{-1}|\,|\underset{\sim}{I} - 2it\,\underset{\sim}{B}|\,|\underset{\sim}{T}| = |\underset{\sim}{I} - 2it\,\underset{\sim}{D}| = \prod_{i=1}^{k} (1 - 2it\lambda_i)^{n_i}. \tag{A.5}$$

Applying the inversion theorem to this characteristic function we find that,

$$|\underset{\sim}{I} - 2it\,\underset{\sim}{A}\,\underset{\sim}{\Lambda}_r|^{-1/2} = \prod_{i=1}^{k} (\,1 - 2it\lambda_i\,)^{-n_i/2}, \tag{A.6}$$

is the characteristic function of $Q_r = \sum \lambda_i Y_i$, where Y_i are independent $\chi^2_{n_i}$ variables. Hence,

$$P_r(\,S^2 \leqq t\,) = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} \Pr(\,Q_r \leqq t\,). \tag{A.7}$$

REFERENCES

[1] Atkins, A.C. and Pearce, M.C., "The computer generation of beta, gamma and normal random variables", Journal of the Royal Statistical Society, A, 139(1976), Part 4, 431-61.

[2] Box, G.E.P., "Nonnormality and tests on variances", Biometrika, 40 (1953), 318-35.

[3] Hyrenius, J., "Distribution of student-Fisher's t in samples from compound normal functions", Biometrika, 37 (1950), 429-42.

[4] Imhof, J.P., "Computing the distribution of quadratic forms in normal variables", Biometrika, 48 (1961), 419-26.

[5] Jensen, D. R. and Solomon, Herbert, "A Gaussian approximation to the distribution of a definite quadratic form", Journal of the American Statistical Association, 67 (1972), 898-902.

[6] Kendall, M.G. and Stuart, A., "The Advanced Theory of Statistics" Vol. I, (1958), Griffin, London.

[7] Koerts, J. and Abrahamese, A.P.J., "On the Theory and Applications of the General Linear Model", (1969) Rotterdam University Press, Rotterdam.

[8] Roy, J. and Tiku, M.L., "A Laguerre series approximation to the sampling distribution of the variance", Sankhya, 24 (1962), 181-4.

[9] Sankaran, Munuswamy, "On the noncentral chisquare distribution", Biometrika, 59 (1972), 235-37.

[10] Subrahmaniam, K., "On quadratic forms from mixtures of two normal populations", South African Statistical Journal, 6 (1972), 103-20.

[11] Tan, W.Y. and Wong, S.P., "On the Roy-Tiku approximation of sample variances from nonnormal universes", Journal of American Statistical Association, 72 (1977), 875-80.

[12] Wilson, Edwin B. and Hilferty Margaret M., "The distribution of chisquare", Proceedings of the National Academy of Sciences, 17 (1931), 684-8.