HONORS PROJECT PAPER

THE ROBUSTNESS OF THE F TEST WITH NON-NORMALITY

JAMES N. WILLIAMS

LYCOMING COLLEGE, WILLIAMSPORT 1989

The Robustness of the F Test with Non-Normality

Presented to the faculty of Lycoming College in partial
fulfillment of the requirements for graduation with
Departmental Honors in Mathematics

Approved by

Gene D. Sprechini

Dr. Gene Sprechini

William E. Keig

Dr. William Keig

Dr. Santhusht deSilva

David G. Fisher

Dr. David Fisher

by

James N. Williams

Lycoming College

April 27, 1989

INTRODUCTION

An analysis of variance (ANOVA) is called "one-way" when one nominal level variable is used to differentiate the populations of interest. This nominal variable is called an independent variable while the variable for which means are to be compared is called a dependent variable. As an example, say we have three brands of wire and we want to know if they all have the same mean breaking strength (in pounds). The brand of wire is the independent variable and the breaking strength is the dependent variable.

The null hypothesis in a one-way ANOVA states that the means for the k populations being compared are equal, while the alternative hypothesis states that there is at least one difference among the k population means. The data is generated by selecting a random sample of items from each population, and can be represented by

$$\begin{array}{llll} x_{11} & x_{12} & \cdots & x_{1n_1} \\ x_{21} & x_{22} & \cdots & x_{2n_2} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ x_{k1} & x_{k2} & \cdots & x_{kn_k} \end{array}$$

where the first subscript indicates from which population an item is taken and the second subscript distinguishes between different items within the same sample.

The means and variances for the k populations are represented respectively by μ_1, μ_2, ..., μ_k and σ_1^2, σ_2^2, ..., σ_k^2. Sample means are denoted $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_k$, and the mean of all $n_*=n_1+n_2+...+n_k$ items taken together is denoted $\bar{x}_*$. The hypotheses can be written as

H_0: $\mu_1=\mu_2=...=\mu_k$

vs

H_1: There is at least one difference among the μ_i's

Returning to the wire example, suppose we have a random sample for each of three brands of wire resulting in the following data.

BRAND α

165 162 159 162

BRAND β

156 163 158

BRAND Γ

151 154 160

We have k=3 samples, the respective sample sizes are $n_1=4$, $n_2=3$, $n_3=3$, and the total sample size is $n_*=10$. The respective sample means are $\bar{x}_1=162$, $\bar{x}_2=159$, and $\bar{x}_3=155$. The mean of all the n_* items taken together is $\bar{x}_*=159$.

The test statistic in a one-way ANOVA is the ratio of two mean squares, one measuring variation between samples and the other measuring variation within samples. The between samples mean square, denoted MSB, is the between samples sum of squares, denoted SSB, divided by its degrees of freedom (k-1), that is,

$$MSB = SSB/(k-1) = \Sigma n_i(\bar{x}_i-\bar{x}_*)^2/(k-1) .$$

Similarly, the within samples mean square sum, denoted MSE, is the within samples sum of squares, denoted SSE, divided by its degrees of freedom (n*-k), that is,

MSE = SSE/(n*-k) = (xij-$\bar{x}$i)2/(n*-k) .

In the wire example,

SSB=(4)(162-159)2+(3)(159-159)2+(3)(155-159)2=84 ,

MSB=84/(3-1)=42 ,

SSE=(165-162)2+(162-162)2+(159-162)2+(162-162)2+
(156-159)2+(163-159)2+(158-159)2+
(151-155)2+(154-155)2+(160-155)2=86 ,

MSE=86/(10-3)=12.286 .

The test statistic is equal to 42/12.286 = 3.42 . As it turns out, the f test with a significance level of 0.05, f(2,7,0.05), yields a value of 4.74, and with a significance level of 0.10, f(2,7,0.10), yields a value of 3.26. Thus the null hypothesis would be rejected under the significance level of 0.05 and accepted under the significance level 0.10 .

The hypothesis test in the one-way ANOVA is based on the fact that the test statistic f(k-1,n*-k) = MSB/MSE has the Fisher's f distribution with (k-1) numerator degrees of freedom and (n*-k) denominator degrees of freedom when the following conditions are satisfied:

(1) Observations are selected independently of one another.

(2) The k populations all have the same variance, i.e. $\sigma_1^2=\sigma_2^2=\ldots=\sigma_k^2$.

(3) The k populations all have normal distributions.

(4) The k populations all have the same mean, i.e. $\mu_1=\mu_2=\ldots=\mu_k$.

Note that condition (4) is the null hypothesis. The null hypothesis is rejected when the test statistic $f(k-1,n_*-k)$ is sufficiently large. After choosing a significance level the f statistic

$$f(k-1,n_*-k) = \frac{SSB/(k-1)}{SSE/(n_*-k)} = \frac{MSB}{MSE}$$

can be compared to an appropriate tabled f value in order to determine whether or not the null hypothesis should be rejected.

Empirical study [1] seems to indicate that the hypothesis test is relatively robust against a variety of departures from the normality condition (3), meaning that the f test "works" reasonably well with many non-normal distributions. However, virtually nothing is known about why this should be the case.

This study is probing into the reason(s) why the f test should work relatively well with many non-normal distributions. One possible explanation could be that non-normal data tend to behave like normally distributed data when put through a suitable orthogonal transformation. The f statistic can be written in an algebraic form were the data are orthogonally transformed. This study utilizes a specific orthogonal transformation of the data based on Helmert matrices [2] to be defined later. It is our contention that the distribution of the orthogonally transformed data will tend to have the appearance of normality.

ONE-WAY ANOVA IN TERMS OF MATRICES

Let **X** be the n_*x1 matrix defined by

$$\mathbf{X}^T=[x_{11},x_{12},\ldots,x_{1n_1},x_{21},x_{22},\ldots,x_{2n_2},\ldots,x_{k1},x_{k2},\ldots,x_{kn_k}].$$

For i=1,2,...,k, let $\mathbf{1}_i$ be the n_ixn_i matrix with every entry equal to one (1); let **M** be the n_*xn_* partitioned matrix with $(1/n_1)\mathbf{1}_1$, $(1/n_2)\mathbf{1}_2$, ..., $(1/n_k)\mathbf{1}_k$ on the diagonal and appropriately dimensioned zero matrices off the diagonal; let $\mathbf{1}_*$ be the n_*xn_* matrix with every entry equal to one (1); let I be the n_*xn_* identity matrix. If **B** and **E** are the n_*xn_* matrices defined by

$$\mathbf{B} = \mathbf{M} - (1/n_*)\mathbf{1}_* \quad \text{and} \quad \mathbf{E} = \mathbf{I} - \mathbf{M} ,$$

then it follows that

$$\text{SSB} = \mathbf{X}^T\mathbf{B}\mathbf{X} \quad \text{and} \quad \text{SSE} = \mathbf{X}^T\mathbf{E}\mathbf{X} \quad .$$

Now, **B** and **E** are each symmetric and idempotent, hence each of their eigenvalues are equal to either 0 or 1. The matrix **B** has rank (k-1), and consequently, there is a (k-1)-dimensional eigenspace corresponding to the eigenvalue 1. The matrix **E** has rank (n_*-k), so there is a (n_*-k)-dimensional eigenspace corresponding to the eigenvalue 1. Since, **BE** = 0, these two eigenspaces will be orthogonal.

Helmert matrices [2] can be used to define an n_*xn_* orthogonal matrix which will simultaneously diagonalize **B** and **E**. Let us denote an mxm Helmert matrix by $\mathbf{H}_m$, that is,

$$H_m = \left[\begin{array}{llllll}
1/\surd(2\cdot 1) & -1/\surd(2\cdot 1) & 0 & 0 \;\ldots & 0 \;\ldots & 0 \\
1/\surd(3\cdot 2) & 1/\surd(3\cdot 2) & -2/\surd(3\cdot 2) & 0 \;\ldots & 0 \;\ldots & 0 \\
 & \vdots & & \vdots & & \\
1/\surd([i+1]\cdot i) & \ldots\ (i \text{ times})\ \ldots & & -i/\surd([i+1]\cdot i)\,..\,0 & \ldots & 0 \\
 & \vdots & & \vdots & & \\
1/\surd(m\cdot[m-1]) & \ldots\ (m-1 \text{ times})\ \ldots & & . & & -(m-1)/\surd(m\cdot[m-1]) \\
1/\surd(m) \;\ldots & . & & . & & \ldots\ 1/\surd(m)
\end{array}\right]$$

Next, we let **H** be the n_*xn_* partitioned matrix with $\mathbf{H}_{n_1}$, $\mathbf{H}_{n_2}$, ... , $\mathbf{H}_{n_k}$ on the diagonal and appropriately dimensional zero matrices off the diagonal. Note that **H** is an orthogonal matrix. Rows n_1, n_1+n_2, $n_1+n_2+n_3$, ..., n_* of the matrix **H** each represent eigenvectors of **E** corresponding to the eigenvalue 0. Hence, the remaining (n_*-k) rows form a basis for the eigenspace corresponding to the eigenvalue 1 of the matrix **E**. An orthogonal basis for the eigenspace corresponding to the eigenvalue 1 of the matrix **B** can be found by selecting (k-1) appropriate linear combinations of rows n_1, n_1+n_2, $n_1+n_2+n_3$, ..., n_* . Let **C** be an n_*xn_* matrix with the property that when **H** is pre-multiplied by **C**, rows n_1, n_1+n_2, $n_1+n_2+n_3$, ..., n_* of **H** are replaced by k orthogonal rows (k-1) of which form an orthogonal basis for the (k-1)-dimensional eigenspace. Let us furthermore require that **C** leave the remaining (n_*-k) rows of **H** fixed. The matrix **P** = **CH** is orthogonal and diagonalizes both **B** and **E**, i.e.

$$\mathbf{B} = \mathbf{P}^T\mathbf{D}_1\mathbf{P} \quad \text{and} \quad \mathbf{E} = \mathbf{P}^T\mathbf{D}_2\mathbf{P}$$

Given that **X** satisfies condition (1), it can be shown [3] that the following are equivalent:

X satisfies condition (3).

The transformed vector **Z=HX** satisfies condition (1).

The transformed vector **Z=HX** satisfies condition (3).

When **X** does not satisfy (3), **Z=HX** will not satisfy either condition (1) or (3). It is conceivable however that **Z=HX** might tend to behave more as if it satisfies conditions (1) and (3) than does **X**. If this were the case, then

$$f(k-1,n_*-k) = \frac{SSB/(k-1)}{SSE/(n_*-k)} = \frac{\mathbf{Z}^T\mathbf{C}^T\mathbf{D}_1\mathbf{C}\mathbf{Z}/(k-1)}{\mathbf{Z}^T\mathbf{C}^T\mathbf{D}_2\mathbf{C}\mathbf{Z}/(n_*-k)}$$

might tend to behave more as if it followed the Fisher's f distribution than one would expect knowing only the distribution of **X**. Hence, the degree to which **Z=HX** departs from conditions (1) and (3) would be at least as instrumental as the degree to which **X** departs from conditions (1) and (3) in determining the robustness of the f test in a one-way ANOVA.

One of the most powerful tests for normality is the Shapiro-Wilk test [4]. Using the Shapiro-Wilk test as a measure of departure from normality, computer simulation can be used to study the behavior of **X** and **Z=HX** for a wide variety of distributions and sample sizes $(n_1,n_2,...n_k)$.

SIMULATION OF RANDOM VARIABLES

All of the distributions studied were generated using independent uniform(0,1) random variables with suitable transformations. Independent uniform(0,1) random variables were generated by scaling random numbers from a computer to be between 0 and 1. It is well known that "random" numbers from a computer are actually pseudorandom (meaning that they will eventually begin to repeat after sufficient time). However, such numbers will pass statistical tests for randomness and for all empirical purposes can be treated as being random. The pseudorandom number generator used [5] was,

$x_i = \mathrm{frac}(\pi + x_{i-1})^5$ where the "seed" x_0 was a value between 0 and 1 , π was set to 3.1415926536, and each x_i was carried out to 10^{-4} .

Tests of randomness were performed with various seeds by generating 10,000 pseudorandom numbers using each particular seed. The number of scaled pseudorandom numbers that were in each of the intervals [0,0.25] , (0.25,0.5] , (0.5,0.75] , (0.75,1] was compared. The more uniformly distributed the scaled pseudorandom numbers are, the closer together the frequency of pseudorandom numbers in the respective intervals should be. The initial seeds tested were arbitrarily chosen. The seed of $x_0 = \pi \times 10^{-1}$ was the one finally chosen for the study.

Independent uniform(0,1) random variables were used to generate a random sample from a normal(0,1) distribution with the Box-Cox transformation. The Box-Cox transformation uses two independent uniform(0,1) random variables U_1 and U_2 to produce two independent normal(0,1) random variables Z_1 and Z_2 as follows:

$$Z_1 = (-2LN(1-U_1))^{1/2} \cdot COS(2U_2\pi)$$

$$Z_2 = (-2LN(1-U_1))^{1/2} \cdot SIN(2U_2\pi)$$

Similarly random variables with other distributions were generated using the following. When U is a uniform(0,1) random variable, then

$X = -\Theta LN(1-U)$ is an exponential distribution with mean equal to Θ ,

$X = (2U-1)^{1/3}$ is a bimodal distribution ,

$X = 1-(1-U)^{1/2}$ is a triangular distribution ,

$X = TAN(\pi(U-0.5))$ is a Cauchy distribution .

Random samples ranging in size from 5 to 50 were generated from each of these distributions. The bimodal distribution used has a pdf

$$f(x) = (3/2)x^2 \quad -1 < x < 1$$

with mean 0 . The triangular distribution has a pdf

$$f(x) = 2-2x \quad 0 < x < 1$$

with mean 1/3. With the Cauchy distribution, the mean does not exist. With the exponential distribution, a mean of $\Theta=1$ was chosen.

RESULTS AND FINDINGS

A computer program was written and encoded to generate the scaled random numbers and transform them into random variables with the different distributions, to compute the necessary Helmert matrix transformations, and to apply the Shapiro-Wilk test. This program was run with total sample sizes (n_*) of 12, 15, 16, 20, 30, 45, 48, and 50 where k ranged between 1 and 10, and each n_i ranged from 2 to 50. In addition, on all the runs the Shapiro-Wilk test was applied not only to the data transformed by the Helmert matrix but also to the untransformed data. Thus the difference in departure from normality could be determined in each case. Each run of the program was done with 1000 iterations, and a counter was used to keep track of the number of times the data was found to be normal using a .05 significance level for the Shapiro-Wilk test. This was the case with both transformed data and untransformed data.

In all of the runs with non-normal distributions the transformed data appeared much more normal than the untransformed data. In fact, with certain non-normal distributions and sample sizes, some of the runs resulted in the transformed data having the appearance of normality almost 90% of the time; while, on the other hand, the untransformed data almost never appeared to be normal.

Out of all the runs there was only one case in which the transformed data did not have a smaller degree of departure from normality. Using a bimodal distribution with a sample size of 50, the untransformed data had the appearance of normality as often as the transformed data when k=1. However, this percentage was 0. (Table 2). Furthermore, when the same data was transformed with k>1, the percentage of times the data appeared to be normal rose considerably higher than 0%. This also seemed to be the case with all the other non-normal data. That is, when transformed with k>1 the appearance of normality occurred more often than when k=1 or, in other words, the degree of departure from normality was lower with the same total sample size when k>1.

The only exception to this was with the Cauchy distribution. The data was transformed with k=1, $n_*=n_i=50$, and also with k=10, $n_i=5$ for each n_i. However, with k=1 the appearance of normality occurred 2.2% of the time, while with k=10 the appearance of normality occurred 2.0% of the time. Perhaps this was the case because of the ratio between k and the sample sizes ($n_i=5$ for each i) being rather large compared to the others. This was the only run where the ratio of k to the sample sizes n_i was larger than one for most n_i.

Other runs were done with the ratio between k and sample size n_i reverse that above. One such case was k=5, $n_i=10$ for each i. This run had the appearance of normality 4.5% of the time with the Cauchy distribution, more then double that of k=10, $n_i=5$ for each i. Even though with k=10 the appearance of normality was less than that for k=1, this was still in support of our contention because they both had percentages that were above that for the untransformed data (1.3%). Furthermore, for all k lower than 10 the percentage went up. (Table 3).

As we have already seen when k>1 the transformed data appear more normal than the transformed data with k=1; however, this rise in the appearance of normality was not proportional to the size of k. In fact, between all the non-normal distributions there was no visible pattern with the size of k and the appearance of normality. It seemed that for each individual non-normal distribution and total sample size (n_*) there was an optimal size of k and optimal sample size(s) (the n_i's) that produced the highest appearance of normality. However, more testing needs to be done in order to see if any pattern does in fact exist. One pattern which did seem to arise with every non-normal distribution except the triangular was that the appearance of normality when $n_i=n_*/k$ for each i was higher than the percentage when $n_i \neq n_*/k$ for each i. (Figures 1,2,3). Of course, both of these percentages were higher than with the untransformed data.

With the normal distribution the transformed data with k>1 was less normal than the untransformed data. (Tables 1,2,3). This would seem to contradict the theorem which states that independence and normality are preserved only with normally distributed data put through orthogonal transformations; however, a possible reason for this drop in normality may be related to the fact that the numbers generated were pseudorandom. The rows n_1, n_1+n_2, $n_1+n_2+n_3$, ...,n_* of the matrix **H** involve the mean of the distribution whereas the other rows do not. With the normal distribution used this mean is 0, but because the numbers are pseudorandom and not truly random the mean may no longer be exactly 0.

CONCLUSION AND SUMMARY

For the most part, the results of the simulations support our hypothesis that orthogonally transformed non-normally distributed data would appear more normal than the untransformed data. Some unexpected patterns arose with the transformed data. In particular, there may exist an optimal size for k and optimal sample sizes (n_i's) yielding the greatest appearance of normality, and that, with the exception of the triangular and normal distributions, when $n_i=n_*/k$ the percentage was higher than when $n_i \neq n_*/k$ for each i.

Furthermore, although the Helmert matrix was employed in this project, the transformation could be based on an infinite number of other orthogonal matrices. This opens the door even further for more research, for a different orthogonal matrix may show different patterns, and possibly there might be a matrix which is "optimal" in some sense, one that is optimal in the sense that it yields the highest appearance of normality for non-normally distributed data. This would give an experimental explanation of "why" the f test is so robust.

SOURCES CITED

[1] Scheffé, H.A., The Analysis of Variance, New York: Wiley (1959), pp. 345-368.

[2] Searle, S.R., Linear Models, New York: Wiley (1971), pp. 32-33.

[3] Anderson, T.W., An Introduction to Multivariate Statistical Analysis, New York: Wiley (1971), pp. 36-37.

[4] Shapiro, S.S. and Wilk, M.B., "An Analysis of Variance Test for Normality", Biometrika, Vol. 52 (1965), pp. 591-611.

[5] Cheney, Ward and Kincaid, David, Numerical Mathematics and Computing, Belmont, California: Wadsworth (1980), pp. 205.

F I G U R E 1

SAMPLE SIZE EQUAL TO 12

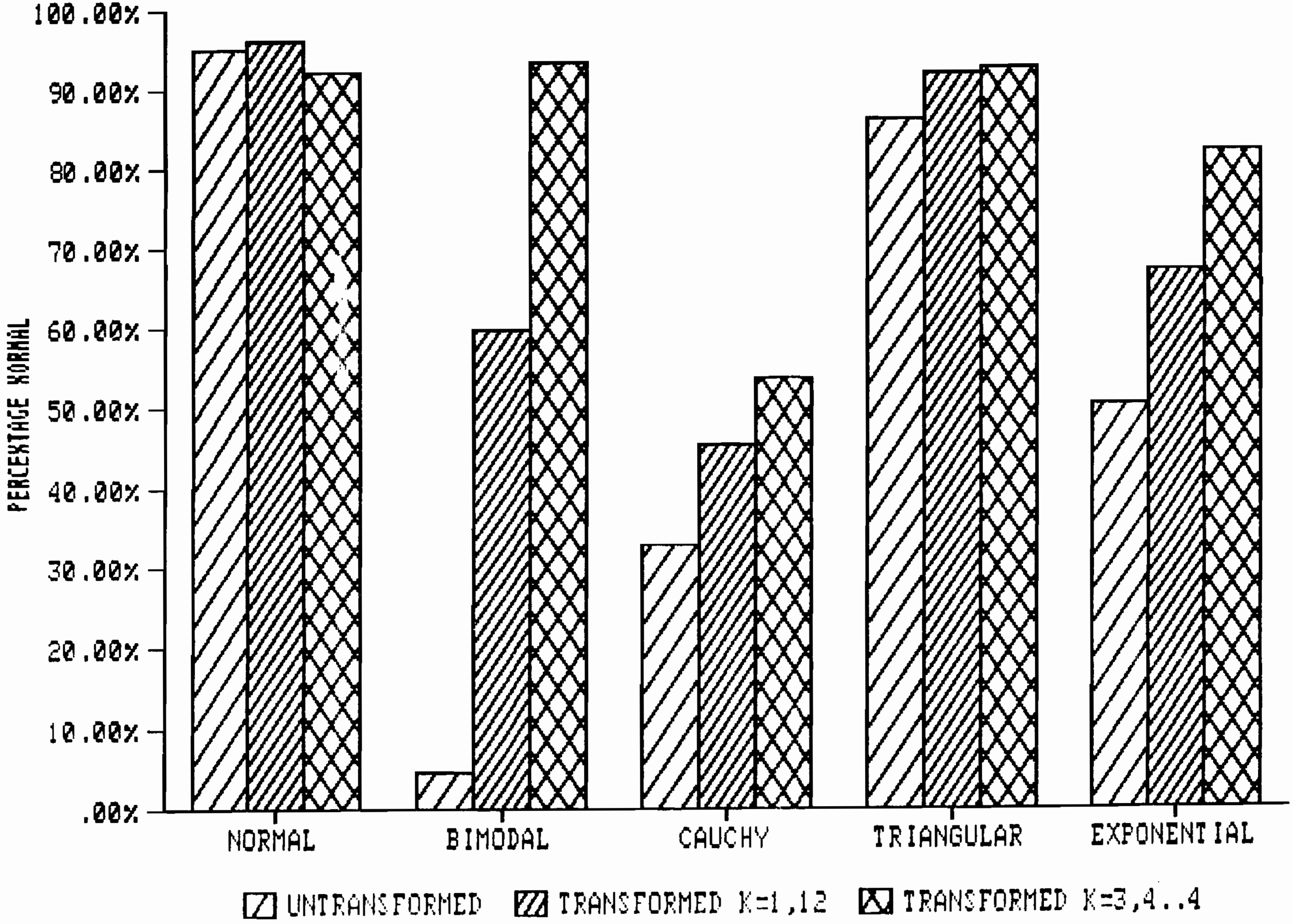

FIGURE 2

SAMPLE SIZE EQUAL TO 30

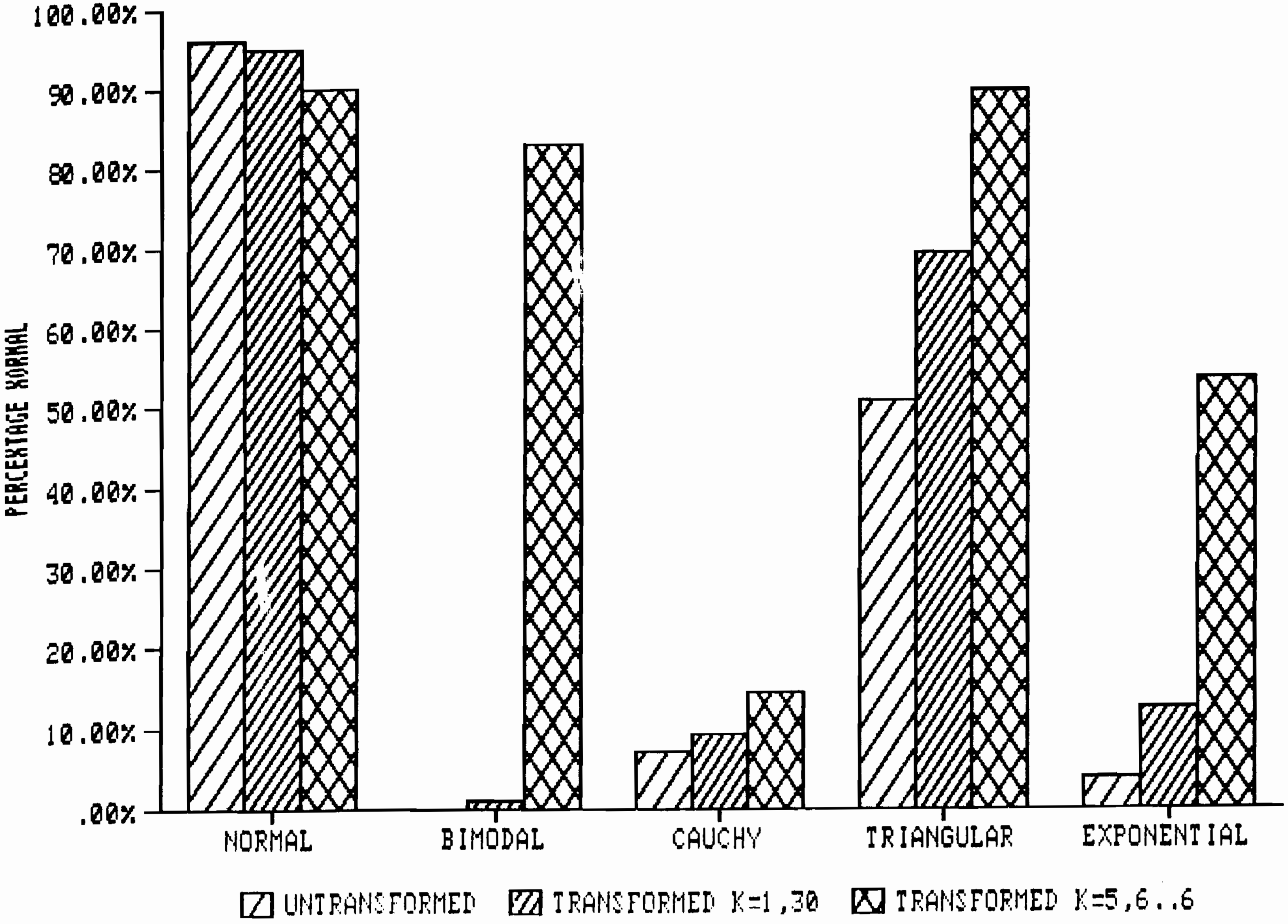

FIGURE 3

SAMPLE SIZE EQUAL TO 50

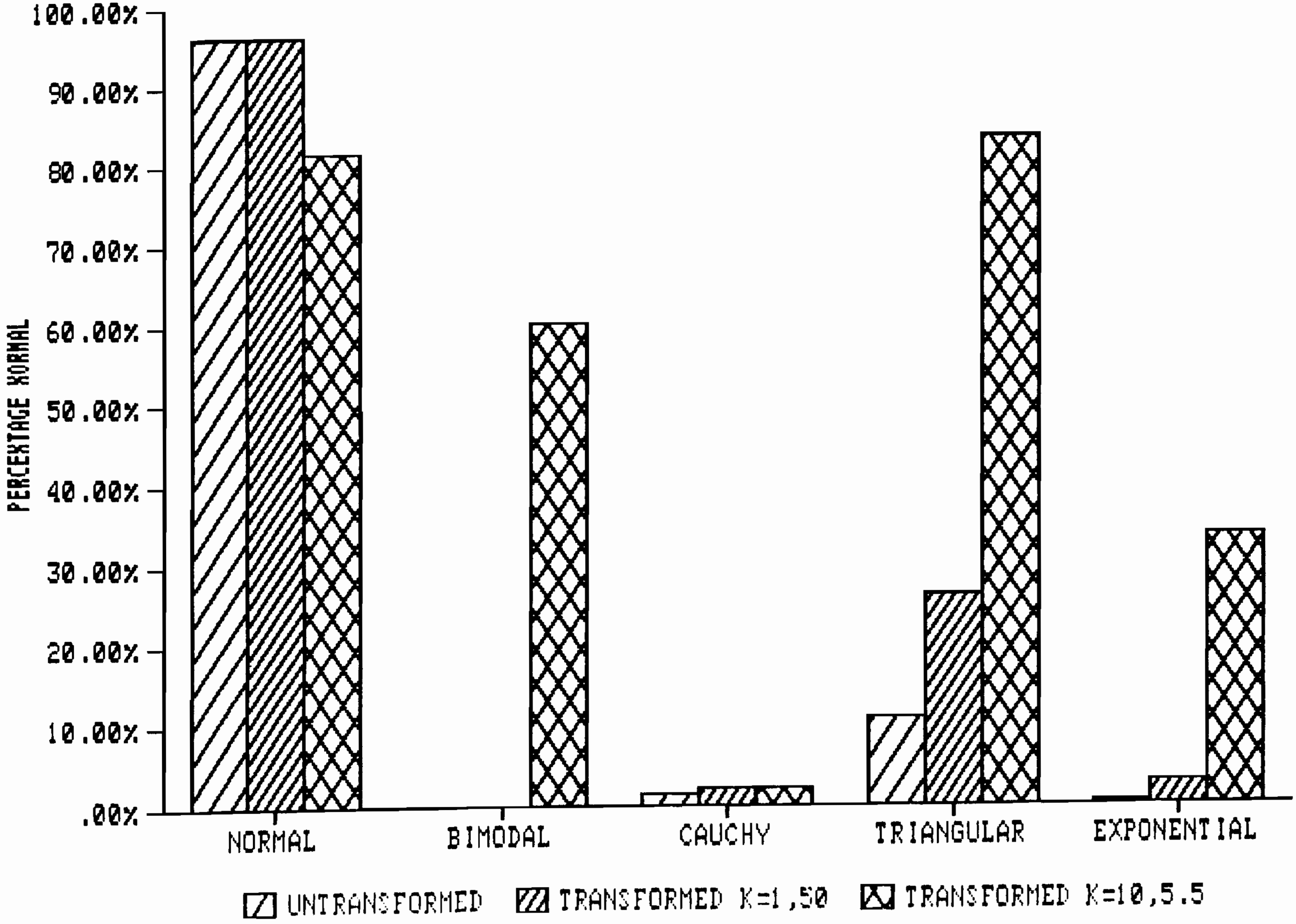

TABLE 3

SEED .314**	NORMAL	BIMODAL	CAUCHY	TRIANGULAR	EXPONENTIAL
UNTRANSFORMED	96.00	0.00	1.30	10.80	0.30
K = 1 50	95.90	0.00	2.20	26.20	2.90
K = 10 5,..,5	81.50	60.00	2.00	83.10	33.70
K = 5 10,..,10	93.90	39.50	4.50	79.00	27.20
K = 5 4,7,10,13,16	92.60	29.20	3.70	76.50	24.60
K = 4 5,10,15,20	94.20	17.70	3.70	71.00	20.10
K = 3 10,15,25	94.60	8.70	3.20	62.40	13.60
K = 2 25,25	95.90	3.60	3.30	49.00	8.80
1000 ITERATIONS	**				

TABLE 2

SEED .314**	NORMAL	BIMODAL	CAUCHY	TRIANGULAR	EXPONENTIAL
UNTRANSFORMED	96.20	0.00	7.00	50.60	3.90
K = 1 30	94.90	0.90	9.30	69.30	12.40
K = 5 6,..,6	90.00	82.80	14.40	89.70	53.60
K = 5 2,4,6,8,10	89.00	80.10	9.80	89.80	48.50
K = 3 10,..,10	93.50	63.20	13.50	86.40	43.30
K = 3 5,10,15	94.40	55.10	12.30	87.60	40.80
1000 ITERATIONS	**				

TABLE 1

SEED .314**	NORMAL	BIMODAL	CAUCHY	TRIANGULAR	EXPONENTIAL
UNTRANSFORMED	94.90	4.80	32.80	86.10	50.30
K = 1 12	96.00	59.80	45.30	91.70	67.20
K = 3 4,..,4	92.00	93.30	53.70	92.50	82.00
K = 3 2,4,6	93.10	90.50	50.90	93.80	80.50
1000 ITERATIONS	**				