Lehigh University
Lehigh Preserve

Theses and Dissertations

1971

Approximation to the double sided distribution of a statistic for testing normality

Charlotte Klever
Lehigh University

Follow this and additional works at: https://preserve.lehigh.edu/etd

 Part of the Operations Research, Systems Engineering and Industrial Engineering Commons

Recommended Citation
Klever, Charlotte, "Approximation to the double sided distribution of a statistic for testing normality" (1971). *Theses and Dissertations*. 3971.
https://preserve.lehigh.edu/etd/3971

This Thesis is brought to you for free and open access by Lehigh Preserve. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Lehigh Preserve. For more information, please contact preserve@lehigh.edu.

APPROXIMATION TO THE DOUBLE SIDED DISTRIBUTION OF A STATISTIC FOR TESTING NORMALITY

by

Charlotte Klever

ABSTRACT

The MB statistic and associated MB test has been proposed for use as a test for normality. Use of the MB test requires knowledge of the distribution of the MB statistics. Originally this distribution was obtained from samples drawn under the null hypothesis of normality by the method of empirical sampling. Expressions are now available which can be used to approximate the null distribution of MB. These approximating expressions are dependent only on the size of the random sample drawn from the population being tested for the property of normality, and can be applied to samples of sizes six to one hundred. The expressions are capable of yielding both a frequency distribution and a cumulative distribution which are extremely close to the empirical distributions of the MB statistics obtained by empirical sampling.

This thesis describes the derivation of the expressions which approximate the MB distributions. Also described in detail is the utilization of the MB test for normality where it is shown how the approximating expressions can be used to calculate critical regions and confidence levels for hypothesis testing.

APPROXIMATION TO THE DOUBLE SIDED DISTRIBUTION OF A STATISTIC FOR TESTING NORMALITY

by

Charlotte Klever

A Thesis

Presented to the Graduate Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Industrial Engineering

Lehigh University

1971

CERTIFICATE OF APPROVAL

This thesis is accepted and approved in partial fulfillment of the requirements for the degree of Master of Science.

May 14, 1971

Date

Sutton Monro

Professor in Charge

Chairman of the Department of
Industrial Engineering

ACKNOWLEDGEMENTS

The author wishes to express her thanks and appreciation to Professor Sutton Monro of Lehigh University for his help and guidance extended during the writing of this thesis. Thanks is also expressed to the other members of the thesis committee Professor Wallace Richardson of Lehigh University and Richard P. Thayer of the Western Electric Engineering Research Center.

TABLE OF CONTENTS

FIGURES

TABLES

TABLES (Cont.)

ABSTRACT

The MB statistic and associated MB test has been proposed for use as a test for normality. Use of the MB test requires knowledge of the distribution of the MB statistics. Originally this distribution was obtained from samples drawn under the null hypothesis of normality by the method of empirical sampling. Expressions are now available which can be used to approximate the null distribution of MB. These approximating expressions are dependent only on the size of the random sample drawn from the population being tested for the property of normality, and can be applied to samples of sizes six to one hundred. The expressions are capable of yielding both a frequency distribution and a cumulative distribution which are extremely close to the empirical distributions of the MB statistics obtained by empirical sampling.

This thesis describes the derivation of the expressions which approximate the MB distributions. Also described in detail is the utilization of the MB test for normality where it is shown how the approximating expressions can be used to calculate critical regions and confidence levels for hypothesis testing.

CHAPTER 1
INTRODUCTION

A statistic, labeled MB, has been proposed and documented for use as a test for normality (1). The statistic is independent of the knowledge of the mean and standard deviation of the population being tested for the property of normality. The value of the MB statistic may be readily calculated from a completely random sample of any size from the unknown population. In order to utilize the MB statistic as a test of normality, the distribution of MB derived under the null hypothesis of normality is required. The null distribution provides the basis for choosing a critical region for the rejection of the alternative hypothesis.

In reference 1, the null distribution of MB was derived by the method of empirical sampling from a normal population, and was then used to analyze the power of the MB test against several types of non-normal populations. The null distribution of MB was obtained by empirical sampling rather than mathematical analysis because of the complexity of the joint distribution of MB.

The method of utilizing the MB test outlined in reference 1 has the drawback of needing to have on hand the percentiles of the empirical distribution of MB, or having to generate the null distribution of MB empirically for each sample size under consideration. The present thesis deals with the development of

approximations to the null and cumulative distributions of MB as functions dependent only upon the size of the sample drawn from the unknown population. Such an approximation to the cumulative distribution of MB would yield the percentile values for the MB statistic, and would therefore, have application in the area of the testing of unknown populations for the property of normality.

CHAPTER 2

MB STATISTIC AND HYPOTHESIS TESTING

2.1 MB Statistic

The MB Statistic for a sample of size n is defined by the equation:

$$MB = -\sum_{i=1}^{n} y_i \ln y_i ,$$

where $$y_i = \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} ,$$

and $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

Under the hypothesis of normality, the distribution of x is:

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}} ,$$ where μ and σ are the unknown population parameters. In order to have application as a successful test for normality, the distribution of MB under the null hypothesis of normality should be well defined and expressible in terms of percentiles as a function of the sample size. In particular, the median of the null distribution of MB for a given sample size was defined in reference one as the MB value of a "perfect" normal sample. A simple relation was then found which expressed MB median as a function of the sample size.

Departures from the property of normality of a population

may be classified as due to either bimodality or long-tailedness (outliers). These departures cause the value of the MB statistic calculated for a sample drawn from the population to depart from the median value of the MB distribution derived under the null hypothesis for the given sample size. The test for normality using MB will have critical regions removed from the median of the null distribution of MB and will involve a double sided test. The double sided MB test is defined such that the farther a sample's MB statistic is from the median value, the greater the probability will be that the sample was drawn from a non-normal population. Values of the MB statistic larger than the MB median for a sample imply bimodality for the population from which the sample was drawn, while values less than the MB median imply outliers or long-tailedness for the population. Table 2-1 gives values of MB obtained by empirical sampling for selected percentage points below and above the median value for several sample sizes.

2.2 Hypothesis Testing

The MB statistic is the basis for the MB test:

Ho: A random sample comes from a normal population against the alternative hypothesis,

Ha: A random sample does not come from a normal population.

Because of the double sided definition of the MB test, several types of test may be performed on the random sample drawn from

an unknown population. The type of test selected depends on the information desired about the population from which the sample is taken. The types of tests are described by the following cases:

Case 1 - Double Sided MB Test

This test is used to establish the critical region for the acceptance of Ho. The steps are:

1. Given an error of the first kind; Pr $\{$ MB is in the critical region if Ho is true $\}$ = α ; look up in Table 2-1 for upper and lower critical values of MB, $MBU_{+\alpha}$ and $MBL_{-\alpha}$.
2. Compute the MB statistic for the random sample of n elements, $x_1, x_2, \ldots, x_n$, using the formula in section 2.1.
3. If the MB value is inside the range $[MBL_{+\alpha}, MBU_{-\alpha}]$, reject Ha in favor of Ho. If the MB value is outside the range, then it is in the critical region and Ho cannot be accepted at this level of α.

For example:

If n = 12 and α = 0.04, then $MBL_{-.04}$ = 1.379 and $MBU_{+.04}$ = 2.203. Therefore, if the MB value of a sample is in the range [1.379, 2.203], the alternative hypothesis Ha can be rejected in favor of Ho at a confidence level of 0.04.

Case 2 - Properties of the Population

It may be determined from the MB statistic calculated for a sample whether the population displays bimodal or long-tailed

characteristics. This is dependent upon whether the MB value calculated for the sample is above or below the median value of the empirical null-distribution for the sample size in question. A test may be defined in which a confidence value may be found on the rejection of Ha: A sample comes from a bimodal type of non-normal population. Likewise, a long-tailed non-normal population could be the subject of Ha. The steps required for this type of test are:

1. Given a random sample of size n with elements $x_1, x_2, \ldots, x_n$, compute the value of the MB statistic using the formula in section 2.1.
2. Using Table 2-1, locate the upper or lower percentile based on the choice of α to determine the type of non-normality and the confidence with which Ha may be rejected.

For example:

i) Bimodal Population.

If n = 12 and the value calculated for MB = 1.920, the percentile found in Table 2-1 is +80%. Since the value calculated for MB is greater than the normal MB median for n = 12, the population from which the sample was drawn has a bimodal characteristic and Ha may be rejected in favor of Ho at a confidence level of 80%.

ii) Long-tailed (Outlier) Population.

If n = 12 and the value of MB calculated for the sample is 1.379,

the percentile found in Table 2-1 is -4%. Since the value calculated for MB is significantly less than normal MB median for n = 12, the population from which the sample was drawn has a definite long-tailed or outlier characteristic and Ha may be rejected in favor of Ho at a low confidence level of 4%.

Case 3 - Single Sided MB Test

If it is desirable to know only if a sample comes from a bimodal (or long-tailed) population, then a single sided MB test may be used. The alternative hypothesis then becomes, Ha: A random sample comes from a bimodal (or long-tailed) type of non-normal population. The steps for this type of test are:

1. Given the error of the first kind, Pr $\{$ MB is in the critical region if Ho is true $\}$ = δ ; look up in Table 2-1 for the critical value of MB. The critical value is defined at $\alpha = 2\delta$ and is MBU $^{+\alpha}$ or MBL $^{-\alpha}$ depending on the alternative hypothesis that is being tested.
2. Compute the value of the MB statistic for the random sample of n elements, $x_1, x_2, \ldots, x_n$, using the formula in section 2.1.
3. If the value of MB that is calculated from the sample is greater (less) than the critical MB value chosen, then Ho cannot be accepted at a confidence level of $\alpha/2 = \delta$. The population from which the sample was drawn may then be said to have bimodal (or long-tailed)

characteristics at the confidence level of $\alpha/2 = \delta$.

If the value of MB calculated from the sample is (less) greater than the critical MB value, then Ha may be rejected in favor of Ho at a confidence level of $\alpha/2 = \delta$.

For example:

If n = 12, δ = 0.02, and the alternative hypothesis chosen is Ha: A random sample comes from a bimodal type of non-normal population, the critical value of MB is at $MBU_{+.04}$ = 2.203. If the value of MB calculated for the sample falls in the region between MB median and the critical MB value, MB less than 2.203, then Ha may be rejected in favor of Ho and the sample can be said as not coming from a bimodal type of non-normal population at a confidence level of 0.02. If the value of MB calculated for the sample is greater than the critical MB value 2.203, then Ho cannot be accepted and the population from which the sample was drawn can be said to exhibit the bimodal characteristic at a confidence level of 0.02.

The single sided MB test is not meaningful for confidence levels greater than δ = 0.50. When δ takes on values greater than 0.50 the critical region includes more than half of the MB null distribution.

At the present time, the tests described in this section can be performed only for those sample sizes and confidence levels tabulated in Table 2-1 unless the time can be taken to

generate the MB null distribution and calculate the percentile values desired. It is the aim of this thesis to allow the quick calculation of MB critical values for all confidence levels and for any sample size; the calculation being dependent only upon the sample size. This will be accomplished by deriving a method of approximating the MB empirical statistics in Table 2-1 which were derived under the null hypothesis of normality.

TABLE 2-1a EMPIRICAL PERCENTAGE POINTS OF MB-STATISTIC FOR NORMAL POPULATION

LOWER VALUES (MBL)

$-\alpha$ Level (based on median with $\alpha = 1$)

N	Median	.98	.90	.80	.70	.60	.50	.40
4	.873	.866	.849	.836	.833	.828	.820	.808
5	1.128	1.118	1.085	1.044	1.001	.960	.922	.880
6	1.281	1.276	1.245	1.210	1.176	1.131	1.081	1.033
7	1.411	1.405	1.381	1.345	1.311	1.273	1.234	1.184
8	1.513	1.505	1.482	1.451	1.424	1.393	1.356	1.314
9	1.619	1.613	1.592	1.556	1.521	1.492	1.458	1.412
10	1.711	1.705	1.683	1.655	1.625	1.590	1.548	1.514
12	1.873	1.866	1.840	1.813	1.784	1.752	1.716	1.683
14	2.008	2.012	1.980	1.956	1.929	1.898	1.870	1.835
16	2.140	2.137	2.114	2.089	2.061	2.029	2.001	1.967
18	2.250	2.247	2.226	2.204	2.175	2.153	2.122	2.084
20	2.343	2.340	2.323	2.302	2.277	2.250	2.218	2.187
25	2.560	2.556	2.536	2.515	2.491	2.470	2.444	2.413
30	2.730	2.728	2.714	2.697	2.677	2.655	2.630	2.603
35	2.879	2.875	2.863	2.848	2.829	2.811	2.786	2.761
40	3.007	3.004	2.991	2.975	2.958	2.941	2.922	2.896
50	3.226	3.224	3.213	3.197	3.182	3.165	3.150	3.127
60	3.406	3.403	3.393	3.379	3.363	3.348	3.332	3.313
70	3.559	3.557	3.546	3.532	3.519	3.506	3.489	3.473
80	3.690	3.687	3.680	3.667	3.654	3.641	3.627	3.611
90	3.806	3.804	3.796	3.786	3.772	3.759	3.748	3.731
100	3.907	3.905	3.898	3.886	3.874	3.862	3.850	3.836

TABLE 2-1a EMPIRICAL PERCENTAGE POINTS OF MB-STATISTIC FOR NORMAL POPULATION

LOWER VALUES (MBL) (Cont'd)

$-\alpha$ Level (based on median with $\alpha = 1$)

N	Median	.30	.20	.12	.08	.04	.02	.01
4	.873	.793	.774	.748	.735	.715	.704	.699
5	1.128	.841	.805	.786	.779	.765	.751	.728
6	1.281	.980	.905	.844	.811	.775	.754	.741
7	1.411	1.121	1.040	.963	.921	.850	.809	.766
8	1.513	1.244	1.166	1.073	1.019	.937	.862	.841
9	1.619	1.363	1.291	1.210	1.133	1.035	.966	.908
10	1.711	1.463	1.399	1.322	1.264	1.154	1.061	.956
12	1.873	1.637	1.578	1.511	1.471	1.379	1.299	1.209
14	2.008	1.787	1.725	1.657	1.601	1.510	1.439	1.379
16	2.140	1.917	1.863	1.805	1.752	1.678	1.578	1.498
18	2.250	2.047	1.991	1.917	1.874	1.802	1.705	1.636
20	2.343	2.141	2.094	2.013	1.974	1.896	1.826	1.797
25	2.560	2.378	2.326	2.277	2.220	2.148	2.099	2.006
30	2.730	2.570	2.523	2.468	2.435	2.360	2.293	2.232
35	2.879	2.723	2.685	2.640	2.594	2.538	2.483	2.419
40	3.007	2.867	2.827	2.776	2.742	2.693	2.643	2.608
50	3.226	3.099	3.065	3.037	3.003	2.963	2.925	2.903
60	3.406	3.294	3.264	3.228	3.199	3.151	3.103	3.061
70	3.559	3.446	3.424	3.392	3.360	3.317	3.298	3.254
80	3.690	3.591	3.564	3.539	3.517	3.482	3.450	3.394
90	3.806	3.713	3.689	3.662	3.646	3.606	3.563	3.544
100	3.907	3.819	3.797	3.776	3.758	3.723	3.693	3.650

TABLE 2-1b EMPIRICAL PERCENTAGE POINTS OF MB-STATISTIC FOR NORMAL POPULATION UPPER VALUES (MBU) (Cont'd)

$+\alpha$ Level (based on Median with $\alpha = 1$)

N	Median	.30	.20	.12	.08	.04	.02	.01
4	.873	1.191	1.253	1.301	1.331	1.359	1.369	1.379
5	1.128	1.372	1.406	1.444	1.461	1.486	1.502	1.509
6	1.281	1.481	1.520	1.576	1.604	1.650	1.679	1.701
7	1.411	1.607	1.654	1.697	1.729	1.768	1.803	1.830
8	1.513	1.723	1.763	1.800	1.828	1.878	1.912	1.937
9	1.619	1.817	1.862	1.901	1.926	1.976	2.013	2.040
10	1.711	1.909	1.945	1.991	2.019	2.052	2.085	2.112
12	1.873	2.062	2.097	2.135	2.161	2.203	2.235	2.277
14	2.008	2.196	2.226	2.275	2.302	2.345	2.375	2.395
16	2.140	2.310	2.347	2.387	2.413	2.452	2.492	2.526
18	2.250	2.410	2.441	2.478	2.501	2.531	2.564	2.594
20	2.343	2.505	2.535	2.569	2.597	2.626	2.664	2.693
25	2.560	2.705	2.734	2.764	2.790	2.816	2.853	2.878
30	2.730	2.875	2.898	2.929	2.950	2.978	3.002	3.048
35	2.879	3.005	3.035	3.069	3.089	3.113	3.143	3.163
40	3.007	3.131	3.154	3.181	3.202	3.228	3.263	3.282
50	3.226	3.336	3.361	3.387	3.408	3.440	3.456	3.476
60	3.406	3.508	3.531	3.554	3.569	3.587	3.615	3.634
70	3.559	3.657	3.675	3.696	3.710	3.730	3.757	3.773
80	3.690	3.778	3.797	3.818	3.834	3.855	3.878	3.896
90	3.806	3.892	3.907	3.927	3.940	3.961	3.978	3.997
100	3.907	3.993	4.010	4.027	4.038	4.057	4.082	4.093

TABLE 2-1b EMPIRICAL PERCENTAGE POINTS OF MB-STATISTIC FOR NORMAL POPULATION

UPPER VALUES (MBU)

$+\alpha$ Level (based on Median with $\alpha = 1$)

N	Median	.98	.90	.80	.70	.60	.50	.40
4	.873	.877	.908	.946	.989	1.039	1.083	1.140
5	1.128	1.136	1.167	1.197	1.236	1.272	1.312	1.347
6	1.281	1.286	1.311	1.341	1.373	1.401	1.428	1.452
7	1.411	1.417	1.439	1.460	1.486	1.515	1.540	1.572
8	1.513	1.519	1.538	1.566	1.590	1.622	1.653	1.689
9	1.619	1.624	1.647	1.673	1.700	1.728	1.756	1.788
10	1.711	1.715	1.742	1.768	1.794	1.819	1.847	1.873
12	1.873	1.878	1.896	1.920	1.946	1.970	2.000	2.034
14	2.008	2.012	2.028	2.059	2.086	2.109	2.135	2.166
16	2.140	2.144	2.161	2.185	2.207	2.226	2.254	2.281
18	2.250	2.255	2.271	2.292	2.313	2.337	2.360	2.382
20	2.343	2.346	2.364	2.387	2.406	2.425	2.452	2.479
25	2.560	2.564	2.578	2.597	2.613	2.636	2.657	2.679
30	2.730	2.733	2.750	2.773	2.791	2.808	2.826	2.848
35	2.879	2.882	2.898	2.916	2.933	2.948	2.965	2.983
40	3.007	3.011	3.025	3.042	3.059	3.075	3.092	3.108
50	3.226	3.229	3.240	3.255	3.272	3.284	3.300	3.318
60	3.406	3.409	3.420	3.432	3.447	3.460	3.474	3.488
70	3.559	3.562	3.571	3.584	3.595	3.609	3.623	3.639
80	3.690	3.692	3.701	3.712	3.724	3.736	3.749	3.763
90	3.806	3.808	3.816	3.827	3.839	3.850	3.862	3.875
100	3.907	3.909	3.917	3.932	3.942	3.955	3.967	3.981

CHAPTER 3
APPROXIMATION TO NULL MB DISTRIBUTION

The steps involved in deriving the approximation to the distribution of the MB statistics under the null hypothesis of normality are discussed in this chapter. The Johnson system of frequency curves (3) was chosen as the system to be used for obtaining the desired approximation.

Step 1: The empirical null distributions of MB for twenty-two sample sizes were generated on an IBM-360/50 computer using the procedures outlined in reference one; with the exception of the random number generator (2). The sample sizes were n=4(1)10, 12(2)20, 25(5)40, and 50(10)100. Two thousand values of MB were generated for each sample size.

Step 2: For each sample size n, the values of β_1 and β_2 were estimated from the empirical data.

$b_1 = m_3^2/m_2^3$ and $b_2 = m_4/m_2^2$, where b_1 is the estimator of β_1, b_2 is the estimator of β_2, and m_i equals the ith moment about the mean of the empirical distribution. The estimates of β_1 and β_2 for each sample size were plotted on the Johnson chart for determining the appropriate distribution approximation. See Figure 3-1. It was concluded from the figure that the Johnson S_B distribution would result in the most reasonable overall fit to the empirical MB null distributions.

The Johnson S_B probability density function, as applied to the MB statistic, is given by the equation

$$f(MB) = \frac{\eta}{\sqrt{2\pi}} \frac{\lambda}{(MB-\epsilon)(\epsilon+\lambda-MB)} \exp\left\{-.5\left[\gamma+\eta \ln\left(\frac{MB-\epsilon}{\epsilon+\lambda-MB}\right)\right]^2\right\}, \quad (3\text{-}1)$$

where $\epsilon \leq MB \leq \epsilon+\lambda$, $\eta>0$, $-\infty<\gamma<\infty$, $\lambda>0$, and $-\infty<\epsilon<\infty$.

Using the Johnson S_B distribution as an approximation to the null MB distribution is equivalent to expressing the null distribution of MB as a transformation of a standard normal variate. Specifically,

$$z = \gamma + \eta \ln\left[\ (MB-\epsilon)\ /\ (\epsilon+\lambda-MB)\ \right], \quad (3\text{-}2)$$

where ϵ is the minimum MB value,

$\epsilon+\lambda$ is the maximum MB value, and

z is a standard normal variate.

Step 3: The Johnson S_B distribution requires knowledge of four parameters which must be derived from the empirical data. These parameters, from equation (3-1), are ϵ, λ, η and γ. The parameter ϵ is a location parameter and is the smallest possible value of the MB statistic. This parameter was set equal to zero. The parameter λ is a scale parameter and represents the range of values of the MB statistic. The value of $\epsilon+\lambda$ is the largest possible value of the MB statistic which is shown in reference 1 to be ln(n).

The remaining parameters η and γ are shape parameters,

and were estimated by the equations

$$\hat{\eta} = \frac{z_{1-\alpha} - z_{\alpha}}{\ln\left[\frac{(MB_{1-\alpha} - \epsilon)(\epsilon + \lambda - MB_{\alpha})}{(MB_{\alpha} - \epsilon)(\epsilon + \lambda - MB_{1-\alpha})}\right]} \qquad (3\text{-}3)$$

and

$$\hat{\gamma} = z_{1-\alpha} - \hat{\eta} \ln\left[(MB_{1-\alpha} - \epsilon)/(\epsilon + \lambda - MB_{1-\alpha})\right] \qquad (3\text{-}4)$$

where z_{α} and $z_{1-\alpha}$ are the α 100th and $(1-\alpha)$100th percentiles of a standard normal distribution, and MB_{α} and $MB_{1-\alpha}$ are the corresponding values from the empirical data.

The parameters $\hat{\eta}$ and $\hat{\gamma}$ were calculated for $\alpha = 0.01(0.01)0.20$, for each sample size for which an empirical null distribution had been generated up to n=50. The results of these calculations are exhibited in Table 3-1, where $\hat{\eta}$ and $\hat{\gamma}$ are shown as functions of α and n.

Step 4: For the values calculated for $\hat{\eta}$ and $\hat{\gamma}$, the value of α was needed which yielded the Johnson S_B distribution that best approximated the empirical MB distribution. The Chi-square goodness of fit test was used to do this. In order to compare the distributions, the range of the empirical MB statistics for each sample size was divided into intervals of 0.05 and the frequency calculated for each cell. The Johnson S_B probability density function, equation (3-1), was then used to calculate the approximated number of observations for each cell with the mid-point of each cell chosen as the value of MB in the equation.

A Chi-square value for $4 \leq n \leq 50$ was calculated using the equation

$$\chi^2 = (\text{MB frequency} - S_B \text{ frequency})^2 / S_B \text{ frequency}$$

for each set of parameters $\hat{\eta}$ and $\hat{\gamma}$. The results of the Chi-square tests are exhibited in Table 3-2. The large values of Chi-square for small sample sizes indicates that the Johnson S_B distribution is unable to yield a good approximation to the MB distribution for very small sample sizes. The relatively low values of Chi-square for the majority of the sample sizes does indicate however, that the Johnson S_B distribution will give a good fit. It can also be seen from Table 3-2 that the quality of fit, as indicated by smallest Chi-square, is dependent only on the value of α as no other relationship is apparent.

Step 5: In spite of the inconsistences of "best fit" found in Step 4, there does exist a smooth and consistent relation of $\hat{\eta}$ and $\hat{\gamma}$ as functions of the sample size for $n > 5$. See Figure 3-2 for $\hat{\eta}$ and $\hat{\gamma}$ vs. n at $\alpha = 0.05$ and $\alpha = 0.15$. The inconsistences of $\hat{\eta}$ and $\hat{\gamma}$ for n=4 and n=5 can be justified because of the poor fit of the Johnson S_B distribution to the MB empirical distribution for these sample sizes.

From the data in Table 3-1 and Figure 3-2 it can be seen that the values of $\hat{\eta}$ and $\hat{\gamma}$ are significantly more dependent on the sample size than on the value of α. Thus, the criterion for selection of the values of $\hat{\eta}$ and $\hat{\gamma}$ was based on the value

of α which yielded the smallest Chi-square value for each sample size. For n=4, n=5 and n=6, the sample sizes for which the Chi-square values were all too large to permit the above basis for the choice of α, an α value of 0.07 was chosen to be used for calculating the values of $\hat{\eta}$ and $\hat{\gamma}$.

The values of $\hat{\eta}$ and $\hat{\gamma}$ for smallest Chi-square are plotted as functions of sample size in Figure 3-3 and listed with the corresponding values of Chi-square and α in Table 3-3. From Figure 3-3, it can be seen that a consistent relationship exists between the parameters $\hat{\eta}$ and $\hat{\gamma}$ and the sample size under the criteria of smallest Chi-square for sample sizes greater than five. Due to the results in this step and Step 4, it was decided to exclude sample sizes four and five from further consideration.

Step 6: Expressions dependent only on the sample size were desired to express the relationships of $\hat{\eta}$ and $\hat{\gamma}$ to the sample size as exhibited in Figure 3-3. Due to the smoothness of the monotonic relations involved for sample sizes greater than five, the method of polynomial regression was chosen to model the functions. A standard computer program, POLYNOMIAL REGRESSION (8), was used to determine the desired expressions. A third degree polynomial as a function of n was found to yield satisfactory approximations to the graphs of Figure 3-3. The regression models being

$$\tilde{\eta} = A_0 + A_1 n + A_2 n^2 + A_3 n^3 , \qquad (3\text{-}5)$$

and $$\tilde{\gamma} = B_0 + B_1 n + B_2 n^2 + B_3 n^3 . \qquad (3\text{-}6)$$

The results of the polynomial regression and the coefficients for the equations are tabulated in Table 3-4.

The work presented so far in this chapter will allow the quick approximation of the MB empirical distribution as a function of sample size for $6 \leq n \leq 50$. Estimates of η and γ can be obtained from the polynomial regression model, equations (3-5) and (3-6) respectively; for a given sample size. These estimates, combined with the minimum value $\epsilon = 0$ and the maximum value $\lambda = \ln(n)$ of the MB statistics, can be substituted into equation (3-1). Then, by selecting values of MB where $0 \leq MB \leq \ln(n)$, an approximation to the MB empirical distribution may be generated from the Johnson S_B density function.

Step 7: The Johnson S_B distribution may also be used to estimate the percentiles of the MB empirical distributions. To find the expected proportion of observations below some value of MB for a given sample size, the values of η and γ can be estimated from the polynomial regression model, equations (3-5) and (3-6) respectively. These estimates are then substituted into equation (3-2), along with the minimum and maximum values of MB for the given sample size, in order to obtain the standard normal variate z. Using a computer subroutine package for finding the area under a

normal curve (7), or a normal distribution table, the proportion of the area under the standard normal curve to the left of z can be calculated. This value then is also the expected proportion of observations below the MB value in question.

In order to determine the accuracy with which the MB empirical percentiles could be approximated by the Johnson S_B equation (3-2), the cumulative distributions were compared for $6 \leq n \leq 50$. The MB empirical cumulative distributions were obtained from Table 2-1. The Johnson S_B cumulative distribution was obtained from equation (3-2) with MB in that equation taking on the values of MB in Table 2-1. Two different Johnson S_B cumulative distributions were generated; the difference being in the method used to generate the parameters η and γ in equation (3-2). The models are as follows:

(1) η and γ were estimated by $\hat{\eta}$ and $\hat{\gamma}$ which are the values of the parameters based on smallest Chi-square and found in Table 3-3.

(2) η and γ were estimated by $\tilde{\eta}$ and $\tilde{\gamma}$ which are the values of the parameters generated by the regression model equations (3-5) and (3-6).

Table 3-5 exhibits the percentile values obtained for n=6, 7, 8, 10, 25 and 50. From this table it can be seen that for n=6 and n=7 the approximations are not exact in the tails, but are satisfactory for most of the cumulative distribution.

For $n \geq 7$, the Johnson S_B distribution does yield a good overall fit to the MB empirical cumulative distribution.

As a means of comparison between the MB empirical cumulative distribution and the approximating Johnson S_B distributions, the sum of squares of the differences was calculated for $6 \leq n \leq 50$. These results are tabulated in Table 3-6. The small sums of squares in the table indicates that a good approximation can be obtained with the Johnson S_B distribution. It can also be seen from Table 3-6 that the method of estimating the parameters η and γ by the regression model compares favorably with the results obtained when $\hat{\eta}$ and $\hat{\gamma}$ were used as the estimators of η and γ.

From Step 7 it can be concluded that the MB percentile values can be approximated as functions of the sample size. This is due to the ability to estimate the parameters η and γ of equation (3-2) by the regression equations (3-5) and (3-6) respectively; these equations being functions of n only for $6 \leq n \leq 50$.

FIGURE 3-1 THE (β_1, β_2) POINTS ESTIMATED FOR THE EMPIRICAL MB DISTRIBUTION

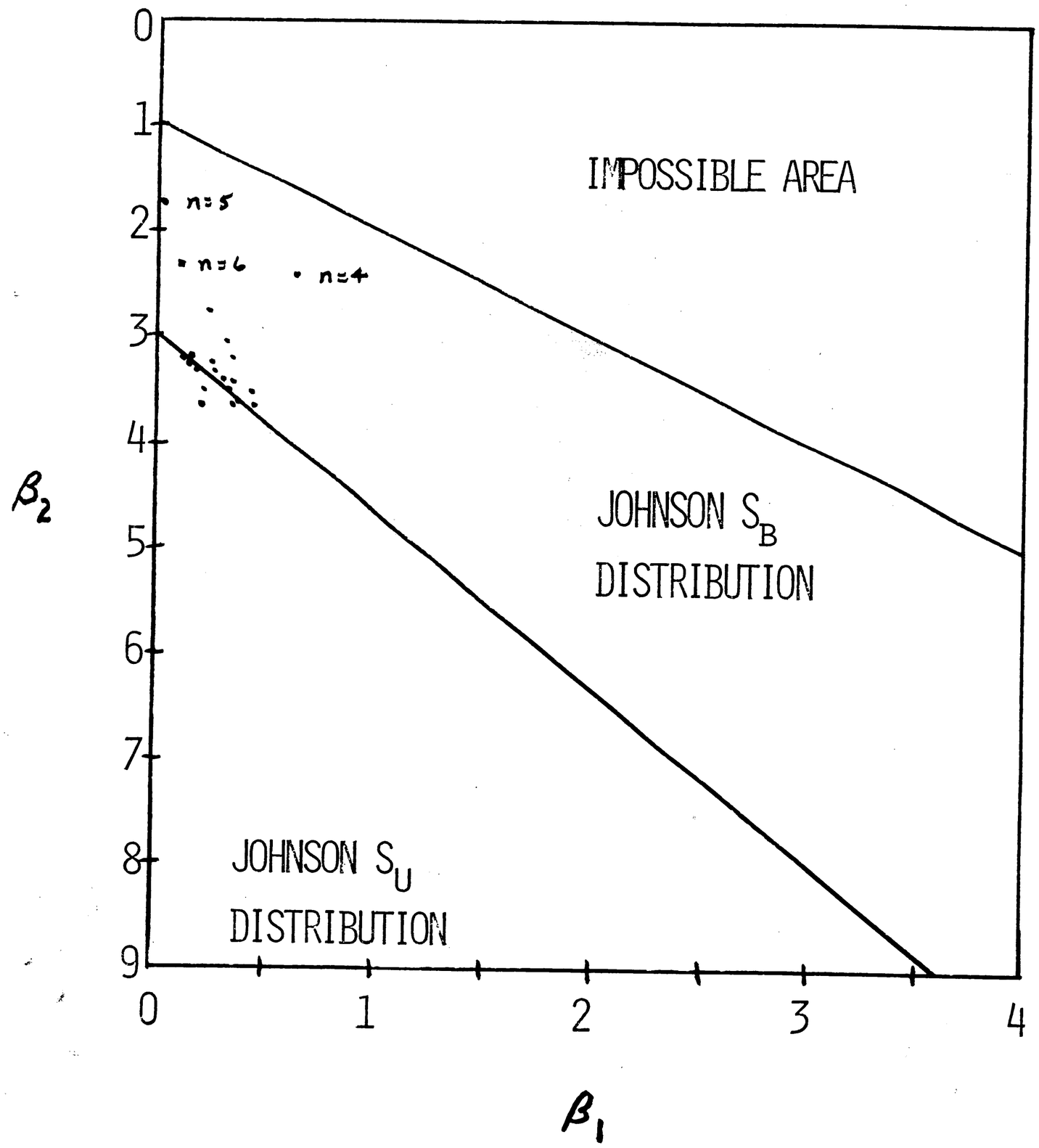

FIGURE 3-2A ESTIMATES OF THE JOHNSON S_B PARAMETERS η AND γ AS FUNCTIONS OF SAMPLE SIZE FOR $\alpha = 0.05$

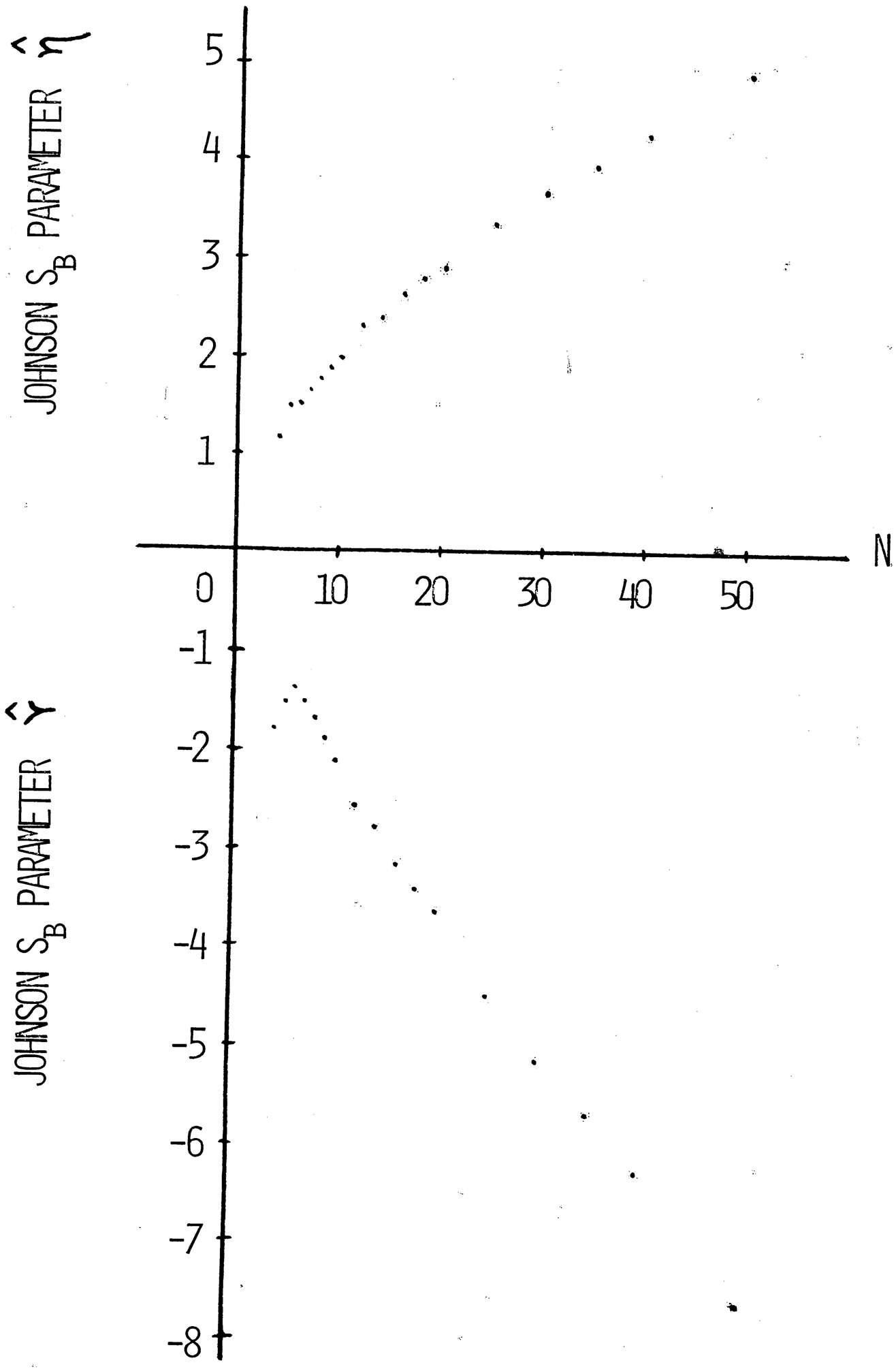

FIGURE 3-2B ESTIMATES OF THE JOHNSON S_B PARAMETERS η AND γ AS FUNCTIONS OF SAMPLE SIZE FOR $\alpha = 0.15$

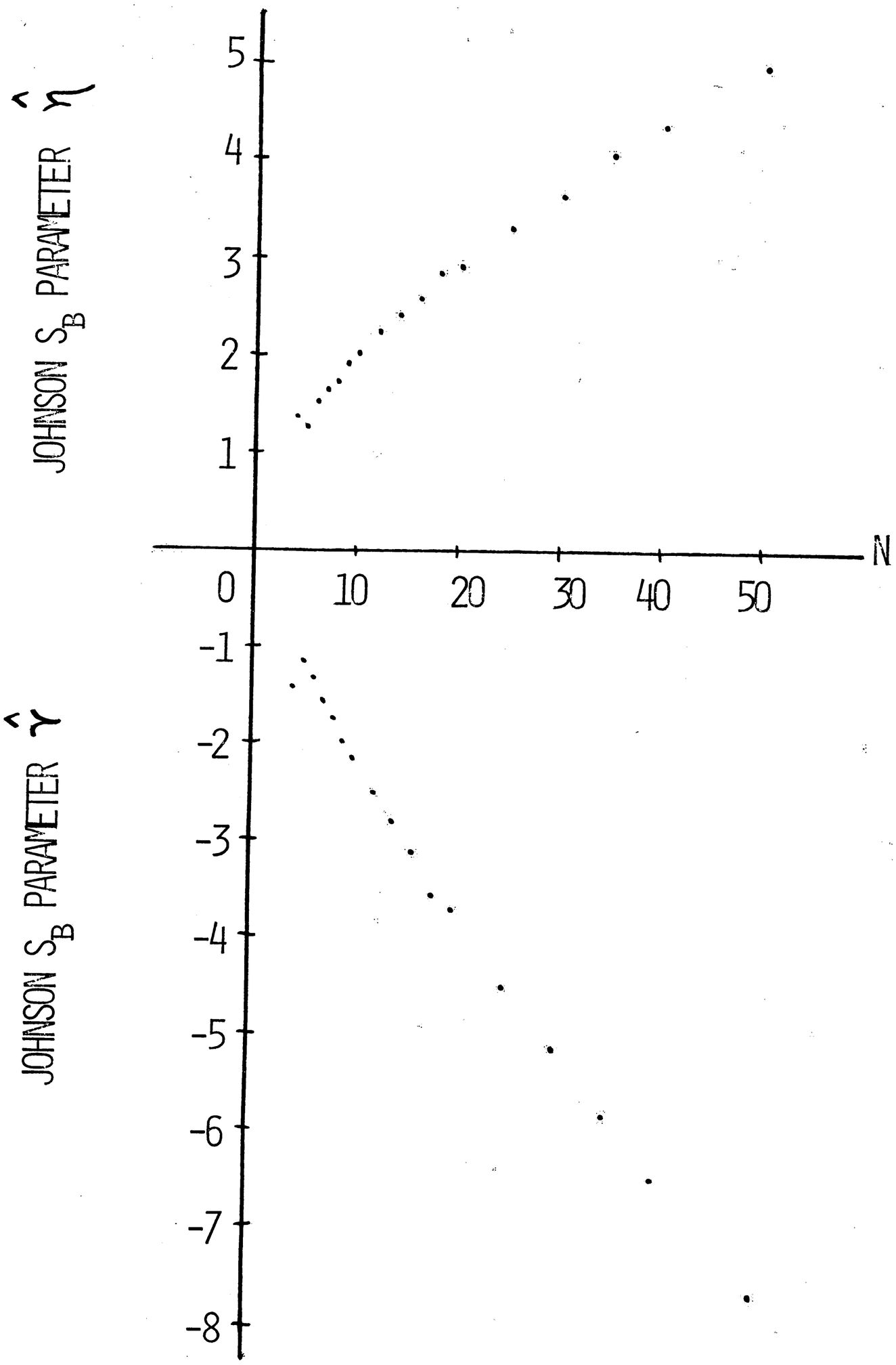

FIGURE 3-3 ESTIMATES OF THE JOHNSON S_B PARAMETERS η AND γ AS FUNCTIONS OF SAMPLE SIZE FOR α YIELDING THE SMALLEST CHI-SQUARE VALUE

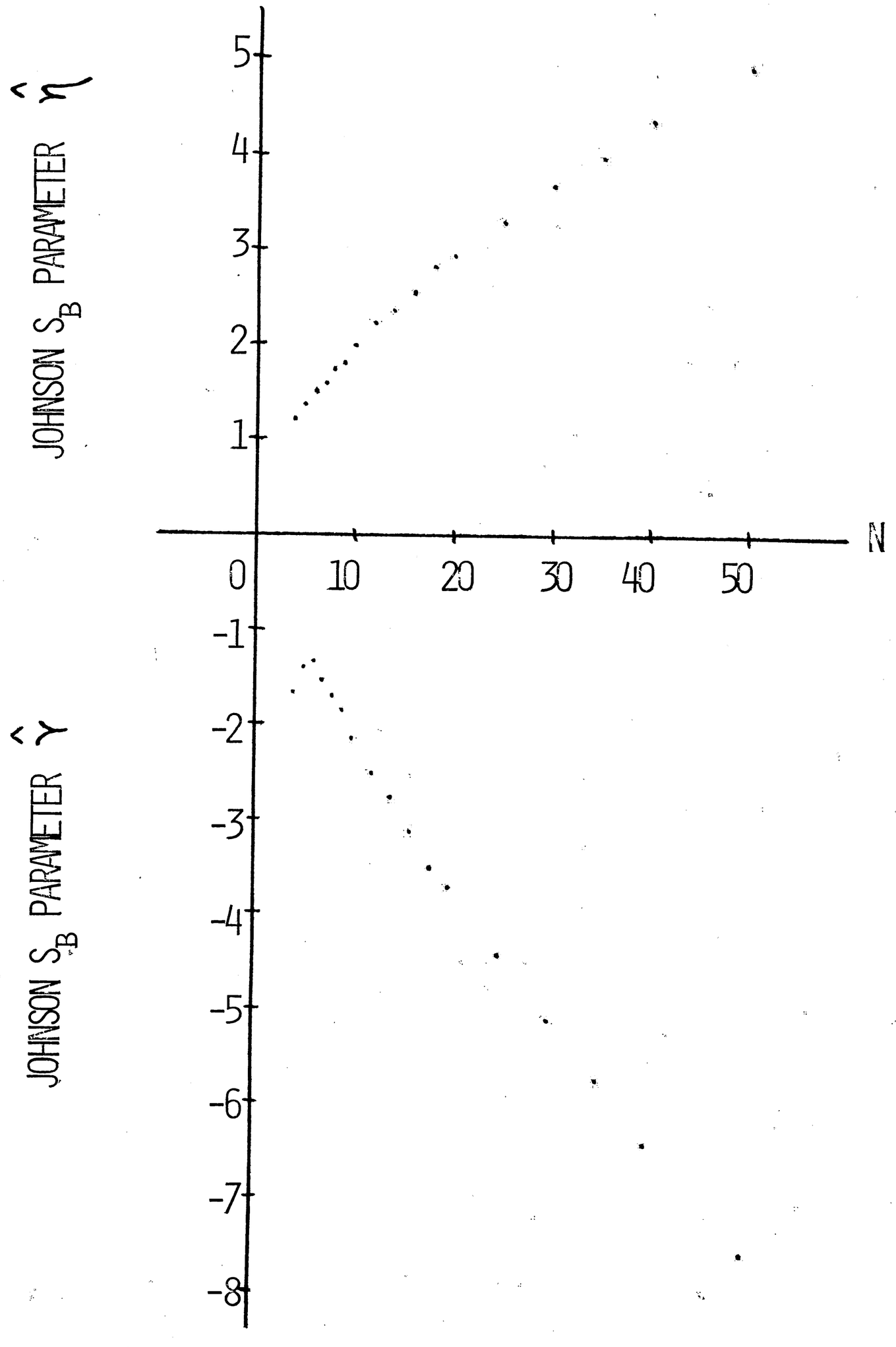

TABLE 3-1a JOHNSON S_B PARAMETER $\hat{\eta}$ CALCULATED FROM MB EMPIRICAL DISTRIBUTION AS A FUNCTION OF SAMPLE SIZE AND α

α	SAMPLE SIZE 4	5	6	7	8	9	10	12	14
.01	1.073	1.678	1.539	1.616	1.674	1.766	1.926	2.216	2.308
.02	1.068	1.589	1.507	1.613	1.691	1.783	1.958	2.236	2.303
.03	1.105	1.548	1.479	1.618	1.717	1.804	1.966	2.295	2.331
.04	1.146	1.492	1.498	1.602	1.729	1.845	1.985	2.297	2.356
.05	1.169	1.445	1.486	1.616	1.741	1.872	1.985	2.290	2.384
.06	1.216	1.406	1.475	1.604	1.730	1.878	2.000	2.271	2.379
.07	1.232	1.380	1.503	1.616	1.754	1.880	2.015	2.287	2.424
.08	1.251	1.346	1.514	1.602	1.749	1.883	1.996	2.281	2.440
.09	1.275	1.328	1.510	1.600	1.740	1.896	2.023	2.265	2.448
.10	1.278	1.362	1.507	1.607	1.740	1.883	2.037	2.260	2.443
.11	1.296	1.303	1.505	1.605	1.764	1.906	2.027	2.256	2.442
.12	1.304	1.289	1.514	1.608	1.763	1.912	2.021	2.252	2.423
.13	1.318	1.269	1.506	1.612	1.748	1.918	2.029	2.260	2.399
.14	1.354	1.261	1.514	1.644	1.752	1.932	2.034	2.253	2.407
.15	1.364	1.245	1.507	1.657	1.761	1.933	2.026	2.238	2.407
.16	1.361	1.232	1.492	1.652	1.783	1.929	2.032	2.263	2.403
.17	1.364	1.204	1.492	1.673	1.784	1.910	2.034	2.243	2.422
.18	1.379	1.192	1.474	1.677	1.777	1.886	2.026	2.225	2.428
.19	1.397	1.175	1.477	1.672	1.768	1.890	2.040	2.214	2.423
.20	1.408	1.163	1.474	1.694	1.824	1.896	2.053	2.204	2.419

TABLE 3-1a JOHNSON S_B PARAMETER $\hat{\eta}$ CALCULATED FROM MB EMPIRICAL DISTRIBUTION AS A FUNCTION OF SAMPLE SIZE AND α (Cont'd)

α	SAMPLE SIZE 16	18	20	25	30	35	40	50
.01	2.442	2.740	2.839	3.265	3.606	3.908	4.193	4.965
.02	2.556	2.838	2.903	3.292	3.625	3.962	4.311	4.844
.03	2.540	2.850	2.886	3.266	3.675	3.923	4.343	4.846
.04	2.570	2.803	2.883	3.256	3.676	3.894	4.268	4.888
.05	2.611	2.780	2.894	3.355	3.680	3.933	4.260	4.897
.06	2.594	2.787	2.887	3.374	3.647	3.979	4.305	5.026
.07	2.583	2.842	2.924	3.359	3.653	3.972	4.346	5.017
.08	2.595	2.847	2.947	3.332	3.677	3.987	4.330	4.954
.09	2.576	2.856	2.959	3.340	3.681	4.043	4.359	4.992
.10	2.587	2.860	2.970	3.321	3.685	4.022	4.369	4.920
.11	2.586	2.838	2.966	3.300	3.674	4.023	4.410	4.905
.12	2.599	2.824	2.942	3.290	3.673	4.038	4.398	4.932
.13	2.620	2.823	2.924	3.311	3.628	4.086	4.377	4.916
.14	2.590	2.841	2.920	3.325	3.639	4.058	4.377	4.986
.15	2.581	2.857	2.910	3.338	3.641	4.042	4.363	4.974
.16	2.618	2.828	2.926	3.359	3.664	4.043	4.386	4.997
.17	2.627	2.814	2.909	3.366	3.626	4.080	4.402	4.959
.18	2.593	2.822	2.927	3.334	3.648	4.108	4.374	4.933
.19	2.593	2.823	2.922	3.354	3.641	4.153	4.433	4.979
.20	2.623	2.830	2.934	3.338	3.692	4.152	4.410	4.976

TABLE 3-1b JOHNSON S_B PARAMETER $\hat{\gamma}$ CALCULATED FROM MB EMPIRICAL DISTRIBUTION AS A FUNCTION OF SAMPLE SIZE AND α

α	SAMPLE SIZE 4	5	6	7	8	9	10	12	14
.01	-2.362	-2.103	-1.837	-1.776	-1.749	-1.898	-2.025	-2.528	-2.746
.02	-2.123	-1.898	-1.644	-1.645	-1.716	-1.848	-2.062	-2.546	-2.724
.03	-1.992	-1.762	-1.531	-1.603	-1.682	-1.862	-2.118	-2.603	-2.758
.04	-1.888	-1.655	-1.466	-1.578	-1.682	-1.866	-2.143	-2.607	-2.772
.05	-1.804	-1.562	-1.412	-1.553	-1.706	-1.908	-2.130	-2.582	-2.789
.06	-1.749	-1.490	-1.382	-1.524	-1.667	-1.938	-2.151	-2.552	-2.801
.07	-1.688	-1.429	-1.356	-1.546	-1.702	-1.933	-2.160	-2.566	-2.859
.08	-1.647	-1.375	-1.301	-1.519	-1.713	-1.940	-2.140	-2.554	-2.853
.09	-1.606	-1.324	-1.323	-1.504	-1.706	-1.961	-2.175	-2.537	-2.842
.10	-1.582	-1.282	-1.314	-1.505	-1.707	-1.949	-2.171	-2.533	-2.833
.11	-1.548	-1.242	-1.311	-1.500	-1.753	-1.964	-2.164	-2.528	-2.841
.12	-1.520	-1.206	-1.317	-1.507	-1.755	-1.971	-2.153	-2.529	-2.826
.13	-1.485	-1.178	-1.306	-1.511	-1.741	-1.973	-2.161	-2.525	-2.789
.14	-1.464	-1.164	-1.307	-1.523	-1.731	-1.992	-2.179	-2.529	-2.817
.15	-1.431	-1.148	-1.320	-1.544	-1.738	-1.987	-2.162	-2.507	-2.819
.16	-1.400	-1.134	-1.308	-1.547	-1.776	-1.980	-2.164	-2.536	-2.812
.17	-1.373	-1.108	-1.310	-1.573	-1.782	-1.963	-2.168	-2.515	-2.836
.18	-1.353	-1.091	-1.294	-1.572	-1.781	-1.936	-2.153	-2.501	-2.843
.19	-1.337	-1.075	-1.291	-1.570	-1.771	-1.942	-2.169	-2.494	-2.839
.20	-1.313	-1.060	-1.297	-1.588	-1.828	-1.956	-2.180	-2.477	-2.839

TABLE 3-1b JOHNSON S_B PARAMETER $\hat{\gamma}$ CALCULATED FROM MB EMPIRICAL DISTRIBUTION AS A FUNCTION OF SAMPLE SIZE AND α (Cont'd)

α	SAMPLE SIZE 16	18	20	25	30	35	40	50
.01	-3.007	-3.325	-3.592	-4.380	-4.948	-5.608	-6.214	-7.723
.02	-3.147	-3.484	-3.636	-4.347	-5.020	-5.676	-6.341	-7.568
.03	-3.106	-3.519	-3.640	-4.340	-5.117	-5.644	-6.416	-7.560
.04	-3.139	-3.465	-3.650	-4.351	-5.148	-5.614	-6.292	-7.591
.05	-3.195	-3.429	-3.647	-4.501	-5.168	-5.708	-6.291	-7.617
.06	-3.173	-3.444	-3.627	-4.534	-5.104	-5.772	-6.346	-7.812
.07	-3.158	-3.531	-3.689	-4.501	-5.124	-5.753	-6.414	-7.771
.08	-3.166	-3.556	-3.721	-4.468	-5.159	-5.773	-6.408	-7.665
.09	-3.144	-3.556	-3.768	-4.468	-5.157	-5.838	-6.443	-7.727
.10	-3.136	-3.558	-3.782	-4.461	-5.173	-5.813	-6.472	-7.609
.11	-3.140	-3.530	-3.771	-4.437	-5.165	-5.800	-6.546	-7.589
.12	-3.154	-3.500	-3.748	-4.439	-5.189	-5.814	-6.536	-7.630
.13	-3.166	-3.507	-3.726	-4.454	-5.123	-5.896	-6.501	-7.603
.14	-3.133	-3.544	-3.720	-4.493	-5.142	-5.861	-6.499	-7.717
.15	-3.117	-3.572	-3.706	-4.508	-5.146	-5.826	-6.489	-7.695
.16	-3.179	-3.543	-3.736	-4.531	-5.184	-5.845	-6.522	-7.738
.17	-3.179	-3.515	-3.702	-4.547	-5.125	-5.904	-6.542	-7.678
.18	-3.140	-3.530	-3.748	-4.504	-5.150	-5.943	-6.499	-7.639
.19	-3.145	-3.528	-3.744	-4.528	-5.136	-6.017	-6.595	-7.716
.20	-3.185	-3.531	-3.761	-4.505	-5.208	-6.020	-6.554	-7.721

TABLE 3-2 CHI-SQUARE VALUES FOR GOODNESS OF FIT TEST BETWEEN JOHNSON S_B DISTRIBUTION AND MB EMPIRICAL DISTRIBUTION

	SAMPLE SIZE								
α	4	5	6	7	8	9	10	12	14
.01	21603.79	2891.01	661.84	154.72	65.13	50.54	24.24	17.08	21.68
.02	13413.70	1804.64	321.44	64.78	49.12	35.63	21.44	15.77	19.31
.03	9905.31	1318.13	212.43	49.38	42.54	33.36	21.11	16.48	17.46
.04	7964.21	1049.94	156.68	47.29	43.18	35.73	20.42	16.98	15.09
.05	6264.19	873.77	139.93	42.77	42.40	35.98	20.36	14.22	14.49
.06	5297.27	766.25	128.52	42.69	45.42	36.10	20.04	12.35	15.38
.07	4681.91	686.66	128.33	42.62	44.61	36.41	20.50	13.32	18.69
.08	4278.92	635.59	133.34	42.81	42.73	37.77	20.56	12.62	18.89
.09	3906.92	586.85	134.62	43.70	42.30	39.18	21.67	11.78	20.90
.10	3766.23	543.45	134.65	44.30	42.28	37.06	22.41	11.64	20.37
.11	3532.07	521.98	134.38	44.66	43.54	41.05	21.20	11.48	19.47
.12	3395.75	502.57	138.62	44.29	45.50	42.54	20.56	11.59	17.00
.13	3238.58	494.53	136.39	44.38	42.66	44.11	21.42	11.54	15.18
.14	3152.23	490.69	140.75	50.13	42.45	47.87	22.37	11.59	15.76
.15	3098.53	490.02	133.78	50.70	43.06	48.46	21.11	11.28	15.79
.16	3052.08	490.83	128.64	48.39	46.01	47.14	21.80	11.71	15.46
.17	3035.51	496.86	128.29	52.14	46.54	42.11	22.01	11.33	16.94
.18	3085.15	496.89	124.30	54.15	46.05	37.30	21.17	11.78	17.59
.19	3199.09	502.34	125.87	52.40	44.87	37.87	22.82	12.46	17.10
.20	3322.35	505.72	123.67	59.24	55.90	39.16	24.91	12.63	16.94

TABLE 3-2 CHI-SQUARE VALUES FOR GOODNESS OF FIT TEST BETWEEN JOHNSON S_B DISTRIBUTION AND MB EMPIRICAL DISTRIBUTION (Cont'd)

	SAMPLE SIZE							
α	16	18	20	25	30	35	40	50
.01	23.88	37.37	20.74	13.78	43.25	17.02	11.41	12.65
.02	18.28	26.47	25.47	27.78	24.53	19.02	16.30	17.01
.03	15.65	23.26	20.55	19.91	18.37	14.27	11.84	14.70
.04	15.77	20.86	18.62	15.68	14.12	13.47	13.33	9.73
.05	19.40	21.88	20.91	13.97	13.60	10.48	11.66	10.78
.06	17.70	20.90	23.07	14.70	13.74	10.25	13.35	13.65
.07	16.77	20.66	20.52	14.62	13.12	10.07	12.76	12.46
.08	17.20	20.79	20.70	13.08	13.62	10.21	10.35	10.09
.09	15.94	21.38	18.69	14.52	14.23	12.34	11.73	11.77
.10	16.33	21.77	19.34	12.33	13.82	11.31	10.88	9.72
.11	16.24	20.30	19.20	11.81	13.36	12.43	12.32	9.50
.12	17.33	20.36	17.92	11.72	14.77	13.88	11.85	9.76
.13	20.10	19.78	17.38	11.86	18.45	15.07	10.85	9.67
.14	16.90	20.33	17.30	12.64	19.10	12.97	10.81	11.70
.15	17.02	21.66	17.20	12.94	19.22	13.21	10.63	10.56
.16	19.30	20.15	17.48	13.73	15.40	12.03	11.69	11.84
.17	20.93	19.16	17.17	14.48	18.82	13.74	12.19	9.96
.18	16.98	19.61	18.00	12.84	13.62	15.82	10.78	9.51
.19	16.84	19.55	18.07	13.63	13.31	20.05	14.19	11.25
.20	19.93	19.77	18.51	12.81	14.86	20.22	12.47	11.33

TABLE 3-3 ESTIMATES OF JOHNSON S_B PARAMETERS BASED ON SMALLEST CHI-SQUARE

n	χ^2	α	$\hat{\eta}$	$\hat{\gamma}$
4	4681.91	.07	1.232	-1.688
5	686.66	.07	1.380	-1.429
6	128.33	.07	1.503	-1.356
7	42.62	.07	1.616	-1.546
8	42.28	.10	1.740	-1.707
9	33.36	.03	1.804	-1.862
10	20.04	.06	2.000	-2.151
12	11.27	.15	2.238	-2.507
14	14.49	.05	2.384	-2.789
16	15.65	.03	2.540	-3.106
18	19.16	.17	2.814	-3.515
20	17.17	.17	2.902	-3.702
25	11.72	.12	3.290	-4.439
30	13.12	.07	3.653	-5.124
35	10.07	.07	3.972	-5.753
40	10.35	.08	4.330	-6.408
50	9.50	.11	4.905	-7.589

TABLE 3-4 ESTIMATES OF JOHNSON S_B PARAMETERS
BASED ON THIRD DEGREE POLYNOMIAL REGRESSION

A. Coefficients

$A_0 = 0.6855260$ $\quad B_0 = -0.8365309 \times 10^{-1}$
$A_1 = 0.1483192$ $\quad B_1 = -0.2248434$
$A_2 = -0.2174305 \times 10^{-2}$ $\quad B_2 = 0.2480604 \times 10^{-2}$
$A_3 = 0.1799766 \times 10^{-4}$ $\quad B_3 = -0.1979058 \times 10^{-4}$

B. Regression Results

n	$\hat{\eta}$	$\tilde{\eta}$	residual	$\hat{\gamma}$	$\tilde{\gamma}$	residual
6	1.503	1.501	.002	-1.356	-1.348	-.008
7	1.616	1.623	-.007	-1.546	-1.543	-.003
8	1.740	1.742	-.002	-1.707	-1.734	.027
9	1.804	1.857	-.053	-1.862	-1.921	.059
10	2.000	1.969	.031	-2.151	-2.104	-.047
12	2.238	2.183	.055	-2.507	-2.459	-.048
14	2.384	2.385	-.001	-2.789	-2.800	.011
16	2.540	2.576	-.036	-3.106	-3.127	.021
18	2.814	2.756	.058	-3.515	-3.443	-.072
20	2.902	2.926	-.024	-3.702	-3.747	.045
25	3.290	3.316	-.026	-4.439	-4.464	.025
30	3.653	3.664	-.011	-5.124	-5.131	.007
35	3.972	3.985	-.013	-5.753	-5.763	.010
40	4.330	4.291	.039	-6.408	-6.375	-.033
50	4.905	4.915	-.010	-7.589	-7.598	.009

C. Regression Model

$$\tilde{\eta} = A_0 + A_1 n + A_2 n^2 + A_3 n^3$$

$$\tilde{\gamma} = B_0 + B_1 n + B_2 n^2 + B_3 n^3$$

TABLE 3-5 PERCENTILE VALUES FOR MB

SAMPLE SIZE = 6

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
.741	.005	.030	.031
.754	.010	.033	.034
.775	.020	.039	.040
.811	.040	.050	.051
.844	.060	.063	.062
.905	.100	.093	.094
.980	.150	.142	.143
1.033	.200	.186	.188
1.081	.250	.234	.236
1.131	.300	.292	.294
1.176	.350	.350	.353
1.210	.400	.400	.402
1.245	.450	.453	.455
1.276	.490	.502	.505
1.281	.500	.510	.512
1.286	.510	.520	.522
1.311	.550	.561	.563
1.341	.600	.611	.613
1.372	.650	.665	.667
1.401	.700	.713	.715
1.428	.750	.758	.760
1.452	.800	.796	.797
1.481	.850	.840	.841
1.520	.900	.892	.892
1.576	.940	.949	.950
1.604	.960	.969	.970
1.750	.980	.990	.990
1.680	.990	.997	.997
1.701	.995	.999	.999

TABLE 3-5 PERCENTILE VALUES FOR MB (Cont'd)

SAMPLE SIZE = 7

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
.776	.005	.014	.014
.809	.010	.018	.018
.850	.020	.025	.025
.921	.040	.043	.043
.963	.060	.057	.058
1.040	.100	.093	.094
1.121	.150	.147	.148
1.184	.200	.202	.204
1.234	.250	.256	.258
1.272	.300	.302	.305
1.311	.350	.354	.357
1.345	.400	.404	.407
1.381	.450	.459	.463
1.405	.490	.498	.502
1.411	.500	.509	.513
1.417	.510	.519	.523
1.439	.550	.556	.560
1.460	.600	.592	.597
1.486	.650	.637	.641
1.515	.700	.686	.690
1.540	.750	.729	.733
1.512	.800	.780	.784
1.607	.850	.834	.838
1.654	.900	.896	.899
1.697	.940	.940	.942
1.729	.960	.965	.966
1.768	.980	.985	.985
1.804	.990	.995	.955
1.830	.995	.998	.998

TABLE 3-5 PERCENTILE VALUES FOR MB (Cont'd)

SAMPLE SIZE = 8

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
.841	.005	.009	.008
.862	.010	.011	.010
.937	.020	.020	.019
1.019	.040	.038	.036
1.073	.060	.055	.052
1.166	.100	.100	.095
1.244	.150	.156	.149
1.314	.200	.222	.214
1.356	.250	.270	.261
1.393	.300	.318	.308
1.424	.350	.360	.351
1.451	.400	.401	.391
1.482	.450	.450	.439
1.505	.490	.488	.478
1.513	.500	.501	.491
1.520	.510	.512	.502
1.538	.550	.543	.533
1.566	.600	.593	.584
1.590	.650	.634	.624
1.622	.700	.690	.681
1.653	.750	.742	.734
1.689	.800	.799	.793
1.723	.850	.850	.844
1.763	.900	.900	.896
1.800	.940	.938	.934
1.828	.960	.960	.958
1.878	.980	.985	.984
1.912	.990	.994	.994
1.937	.995	.998	.998

TABLE 3-5 PERCENTILE VALUES FOR MB (Cont'd)

SAMPLE SIZE = 10

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
.956	.005	.002	.003
1.061	.010	.007	.008
1.154	.020	.016	.018
1.264	.040	.039	.043
1.322	.060	.060	.065
1.399	.100	.101	.107
1.463	.150	.149	.156
1.514	.200	.198	.206
1.548	.250	.238	.246
1.590	.300	.293	.301
1.625	.350	.344	.351
1.655	.400	.391	.398
1.683	.450	.439	.445
1.705	.490	.479	.485
1.711	.500	.490	.496
1.715	.510	.497	.502
1.742	.550	.546	.551
1.768	.600	.594	.598
1.794	.650	.644	.648
1.819	.700	.690	.693
1.847	.750	.741	.742
1.873	.800	.786	.787
1.909	.850	.843	.843
1.945	.900	.892	.891
1.990	.940	.940	.939
2.019	.960	.962	.961
2.052	.980	.980	.979
2.085	.990	.991	.991
2.112	.995	.996	.996

TABLE 3-5 PERCENTILE VALUES FOR MB (Cont'd)

SAMPLE SIZE = 25

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
2.006	.005	.003	.003
2.099	.010	.009	.009
2.148	.020	.016	.016
2.220	.040	.035	.035
2.277	.060	.062	.062
2.326	.100	.098	.098
2.378	.150	.154	.155
2.413	.200	.203	.204
2.444	.250	.255	.257
2.470	.300	.304	.306
2.491	.350	.348	.350
2.514	.400	.400	.404
2.536	.450	.452	.456
2.556	.490	.499	.504
2.560	.500	.510	.514
2.564	.510	.520	.524
2.578	.550	.556	.560
2.597	.600	.604	.609
2.613	.650	.645	.650
2.636	.700	.701	.706
2.657	.750	.748	.753
2.679	.800	.797	.802
2.705	.850	.847	.852
2.734	.900	.894	.898
2.764	.940	.933	.936
2.790	.960	.958	.960
2.816	.980	.975	.976
2.853	.990	.990	.990
2.878	.995	.995	.995

TABLE 3-5 PERCENTILE VALUES FOR MB (Cont'd)

SAMPLE SIZE = 50

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
2.903	.005	.008	.008
2.925	.010	.012	.012
2.963	.020	.022	.022
3.003	.040	.042	.042
3.037	.060	.069	.069
3.065	.100	.100	.101
3.099	.150	.153	.154
3.127	.200	.209	.211
3.150	.250	.266	.268
3.165	.300	.307	.309
3.182	.350	.356	.358
3.197	.400	.403	.406
3.213	.450	.456	.458
3.224	.490	.493	.496
3.226	.500	.501	.504
3.229	.510	.512	.515
3.240	.550	.548	.552
3.255	.600	.604	.607
3.272	.650	.660	.662
3.284	.700	.700	.703
3.300	.750	.751	.754
3.318	.800	.803	.805
3.336	.850	.846	.849
3.361	.900	.899	.901
3.387	.940	.940	.941
3.408	.960	.963	.964
3.440	.980	.984	.985
3.456	.990	.990	.991
3.476	.995	.995	.995

TABLE 3-6 SUM OF SQUARES OF DIFFERENCE BETWEEN CUMULATIVE DISTRIBUTION OF MB AND JOHNSON S_B

n	JOHNSON S_B $\hat{\eta}$ AND $\hat{\gamma}$	JOHNSON S_B $\tilde{\eta}$ AND $\tilde{\gamma}$
6	.0038	.0041
7	.0022	.0021
8	.0019	.0029
9	.0019	.0008
10	.0014	.0007
12	.0012	.0026
14	.0018	.0029
16	.0020	.0005
18	.0007	.0011
20	.0008	.0006
25	.0006	.0011
30	.0019	.0029
35	.0012	.0017
40	.0027	.0007
50	.0007	.0012

CHAPTER 4

APPROXIMATION FOR LARGE SAMPLE SIZES

The work presented in Chapter 3 is valid for sample sizes $6 \leq n \leq 50$. The work presented in this chapter will deal with large sample sizes where $50 < n \leq 100$. The empirical MB null distributions for sample sizes n=50(10)100 are tabulated in Table 2-1, and it is these null distributions that were used to obtain an approximation for large n. The Johnson S_B distribution, equation 3-1, was also used to approximate large n. The procedures used are the same as the procedure outlined in Chapter 3.

The Johnson S_B parameters η and γ were estimated by equations (3-3) and (3-4) respectively, for each $\alpha = 0.01(0.01)0.02$. A Chi-square value was then calculated for the difference between the Johnson S_B distribution generated for each set of parameters $\hat{\eta}$ and $\hat{\gamma}$ and the MB null distribution. Using the smallest Chi-square criterian, the choice of the values of $\hat{\eta}$ and $\hat{\gamma}$ for each sample size was based on the value of α which yielded the smallest Chi-square for each sample size. The values of $\hat{\eta}$ and $\hat{\gamma}$ chosen are exhibited in Table 4-1 with the corresponding values of Chi-square and α.

In order to obtain expressions for the parameters $\hat{\eta}$ and $\hat{\gamma}$ as functions of n, the values of these parameters for small n taken from Table 3-3 were combined with the values in Table 4-1.

The graphs of the parameters $\hat{\eta}$ and $\hat{\gamma}$ as functions of n for $6 \leq n \leq 100$ are shown in Figure 4-1. It can be seen from the graphs that a smooth relationship still exists even when the values for large n are included.

Polynomial regression was again used to determine the desired expressions, with the third degree polynomial as a function of n yielding a satisfactory approximation to the graphs in Figure 4-1. The results obtained from the regression models

$$\tilde{\eta} = C_0 + C_1 n + C_2 n^2 + C_3 n^3 \qquad (4\text{-}1)$$

and $$\tilde{\gamma} = D_0 + D_1 n + D_2 n^2 + D_3 n^3 \qquad (4\text{-}2)$$

for $6 \leq n \leq 100$ and the coefficients for the equations are exhibited in Table 4-2. Comparing Table 3-4 and Table 4-2, the residuals for $6 \leq n \leq 50$ are greater when large n is included in the model. Therefore, the equations (4-1) and (4-2) are not expected to give as good a result as equations (3-5) and (3-6) for $6 \leq n \leq 50$ and so will be used only for sample sizes where $50 < n \leq 100$.

Using equations (4-1) and (4-2) for the estimates of η and γ respectively in equation (3-1), approximations to the empirical MB distributions may be generated for large sample sizes. The approximations, as for small sample sizes, will be dependent only on the sample size which has been drawn from the unknown population.

The MB empirical percentile values may also be approximated for large sample sizes from equation (3-2). The values of η

and γ in equation (3-2) may be estimated by the regression equations (4-1) and (4-2). In order to determine the accuracy with which the MB empirical percentiles could be approximated by the Johnson S_B equation (3-2), the cumulative distributions were compared for n = 60, 70, 80, 90, and 100. The MB empirical cumulative distributions were obtained from Table 2-1. The Johnson S_B cumulative distributions were obtained from equation (3-2) with MB in that equation taking on the values of MB in Table 2-1. Two different Johnson S_B cumulative distributions were generated for each sample size as follows:

(1) η and γ were estimated by $\hat{\eta}$ and $\hat{\gamma}$ from Table 4-1

(2) η and γ were estimated by $\tilde{\eta}$ and $\tilde{\gamma}$ found by equations (4-1) and (4-2) respectively.

Table 4-3 exhibits the percentile values obtained for n = 60, 80, and 100. It can be seen from the table that the Johnson S_B cumulative distribution does yield a good approximation for large sample sizes. As a means of comparison between the distributions the sum of squares of the differences between the empirical distribution and the Johnson S_B distribution was calculated for the two Johnson S_B models. These values are tabulated in Table 4-4. From Table 4-4 it can be seen that the method of estimating the parameters η and γ by the regression model equations (4-1) and (4-2) respectively, compares favorably with the results obtained when $\hat{\eta}$ and $\hat{\gamma}$ were used as estimators of η and γ in equation (3-2). Therefore, it can be concluded that the

MB percentile values for large n can be approximated as functions of sample size only, and are valid for sample sizes $50 < n \leq 100$ when the regression coefficients in Table 4-2 are used to estimate η and γ.

FIGURE 4-1A ESTIMATE OF THE JOHNSON S_B PARAMETER η AS A FUNCTION OF SAMPLE SIZE - INCLUDING LARGE N

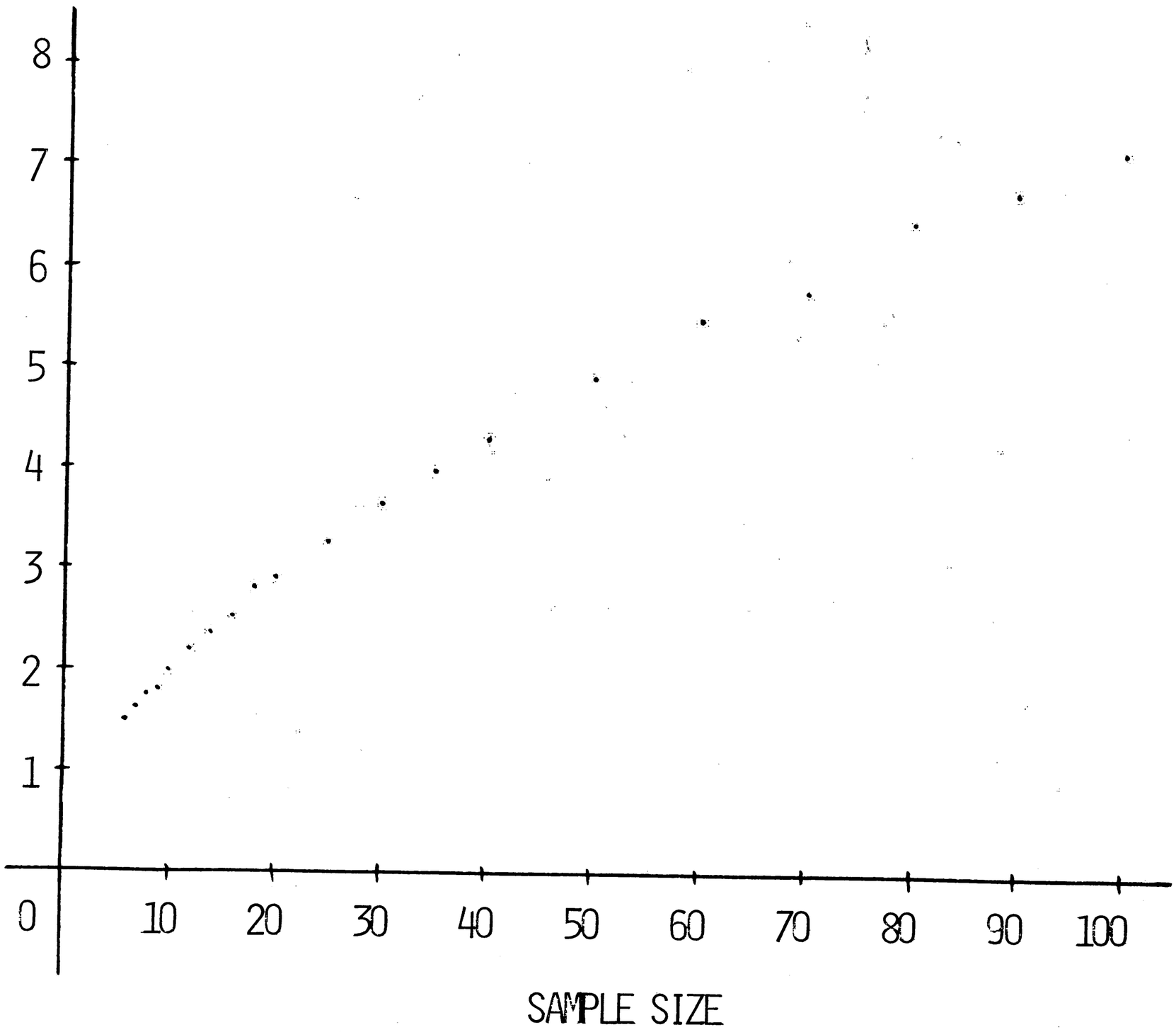

FIGURE 4-1B ESTIMATE OF THE JOHNSON S_B PARAMETER γ AS A FUNCTION OF SAMPLE SIZE - INCLUDING LARGE N

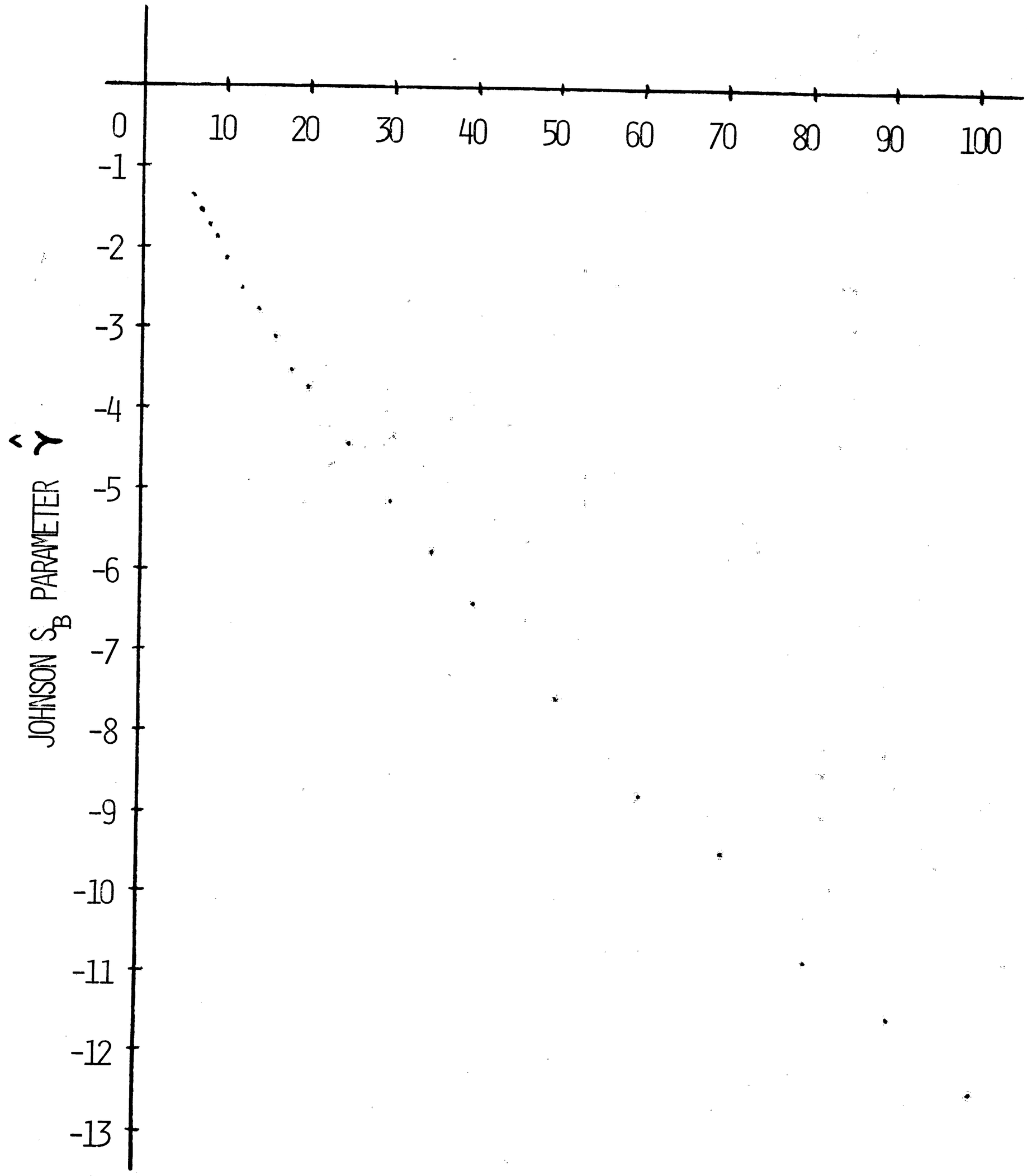

TABLE 4-1 ESTIMATES OF JOHNSON S_B PARAMETERS FOR LARGE n BASED ON SMALLEST CHI-SQUARE

n	χ^2	α	$\hat{\eta}$	$\hat{\gamma}$
60	12.06	.06	5.470	-8.752
70	12.04	.12	5.786	-9.482
80	4.80	.04	6.453	-10.805
90	7.94	.17	6.761	-11.513
100	13.55	.07	7.194	-12.442

TABLE 4-2 ESTIMATES OF JOHNSON S_B PARAMETERS INCLUDING LARGE n
BASED ON THIRD DEGREE POLYNOMIAL

A. Coefficients

$C_0 = 0.8727656$
$C_1 = 0.1180268$
$C_2 = -0.9350425 \times 10^{-3}$
$C_3 = 0.3898801 \times 10^{-5}$

$D_0 = -0.2855893$
$D_1 = -0.1919572$
$D_2 = 0.1123195 \times 10^{-2}$
$D_3 = -0.4219334 \times 10^{-5}$

B. Regression Results

n	$\hat{\eta}$	$\tilde{\eta}$	Residual	$\hat{\gamma}$	$\tilde{\gamma}$	Residual
6	1.503	1.548	-.045	-1.356	-1.398	.042
7	1.616	1.654	-.038	-1.546	-1.576	.030
8	1.740	1.759	-.019	-1.707	-1.752	.045
9	1.804	1.862	-.058	-1.862	-1.925	.063
10	2.000	1.963	.037	-2.151	-2.097	-.054
12	2.238	2.161	.077	-2.507	-2.435	-.072
14	2.384	2.353	.031	-2.789	-2.764	-.025
16	2.540	2.538	.002	-3.106	-3.087	-.019
18	2.814	2.717	.097	-3.515	-3.402	-.113
20	2.902	2.890	.012	-3.702	-3.709	.007
25	3.290	3.300	-.010	-4.439	-4.448	.009
30	3.653	3.677	-.024	-5.124	-5.147	.023
35	3.972	4.025	-.053	-5.753	-5.809	.056
40	4.330	4.347	-.017	-6.408	-6.437	.029
50	4.905	4.924	-.019	-7.589	-7.603	.014
60	5.470	5.430	.040	-8.752	-8.671	-.081
70	5.786	5.890	-.104	-9.482	-9.666	.184
80	6.453	6.327	.126	-10.805	-10.614	-.191
90	6.761	6.764	-.003	-11.513	-11.540	.027
100	7.194	7.224	-.030	-12.442	-12.469	.027

C. Regression Models

$$\tilde{\eta} = C_0 + C_1 n + C_2 n^2 + C_3 n^3$$
$$\tilde{\gamma} = D_0 + D_1 n + D_2 n^2 + D_3 n^3$$

TABLE 4-3 PERCENTILE VALUES FOR MB FOR LARGE n

SAMPLE SIZE = 60

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
3.061	.005	.002	.003
3.103	.010	.006	.007
3.151	.020	.016	.017
3.199	.040	.037	.039
3.228	.060	.060	.063
3.264	.100	.104	.109
3.294	.150	.155	.161
3.313	.200	.198	.204
3.332	.250	.248	.255
3.348	.300	.295	.302
3.363	.350	.342	.349
3.379	.400	.397	.404
3.393	.450	.449	.456
3.403	.490	.488	.495
3.406	.500	.498	.505
3.409	.510	.511	.518
3.420	.550	.553	.560
3.432	.600	.597	.603
3.447	.650	.655	.660
3.460	.700	.703	.707
3.474	.750	.750	.754
3.488	.800	.794	.797
3.508	.850	.850	.852
3.531	.900	.901	.902
3.554	.940	.940	.941
3.569	.960	.958	.958
3.587	.980	.974	.975
3.615	.990	.989	.989
3.634	.995	.995	.995

TABLE 4-3 PERCENTILE VALUES FOR MB FOR LARGE n (Cont'd)

SAMPLE SIZE = 80

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
3.394	.005	.002	.002
3.450	.010	.009	.010
3.482	.020	.019	.020
3.517	.040	.040	.041
3.539	.060	.060	.062
3.564	.100	.095	.096
3.591	.150	.148	.148
3.611	.200	.199	.198
3.627	.250	.250	.248
3.641	.300	.299	.295
3.654	.350	.348	.343
3.667	.400	.398	.392
3.680	.450	.453	.446
3.687	.490	.486	.478
3.690	.500	.497	.489
3.692	.510	.506	.498
3.701	.550	.546	.537
3.712	.600	.598	.588
3.724	.650	.647	.637
3.736	.700	.699	.688
3.749	.750	.750	.740
3.763	.800	.798	.788
3.778	.850	.847	.837
3.797	.900	.896	.888
3.818	.940	.938	.932
3.834	.960	.960	.955
3.855	.980	.979	.976
3.878	.990	.991	.989
3.896	.995	.996	.995

TABLE 4-3 PERCENTILE VALUES FOR MB FOR LARGE n (Cont'd)

SAMPLE SIZE = 100

MB	MB EMPIRICAL PERCENTILE	S_B PERCENTILE $\hat{\eta}$ AND $\hat{\gamma}$	S_B PERCENTILE $\tilde{\eta}$ AND $\tilde{\gamma}$
3.650	.005	.002	.003
3.693	.010	.008	.009
3.723	.020	.019	.020
3.758	.040	.042	.044
3.776	.060	.062	.064
3.797	.100	.096	.099
3.819	.150	.142	.146
3.836	.200	.187	.193
3.850	.250	.236	.243
3.862	.300	.279	.287
3.874	.350	.329	.338
3.886	.400	.381	.390
3.898	.450	.434	.443
3.905	.490	.469	.479
3.907	.500	.477	.486
3.909	.510	.489	.498
3.917	.550	.528	.538
3.932	.600	.599	.609
3.942	.650	.650	.660
3.955	.700	.709	.718
3.967	.750	.760	.769
3.981	.800	.812	.820
3.993	.850	.853	.859
4.010	.900	.900	.905
4.026	.940	.935	.939
4.038	.960	.953	.956
4.057	.980	.975	.976
4.082	.990	.990	.991
4.093	.995	.994	.995

TABLE 4-4 SUM OF SQUARES OF DIFFERENCE BETWEEN CUMULATIVE DISTRIBUTION OF MB AND JOHNSON S_B

n	JOHNSON S_B $\hat{\eta}$ AND $\hat{\gamma}$	JOHNSON S_B $\tilde{\eta}$ AND $\tilde{\gamma}$
60	.0003	.0007
70	.0006	.0003
80	.0001	.0018
90	.0004	.0017
100	.0042	.0026

CHAPTER 5

UTILIZATION OF THE APPROXIMATION FOR MB

The approximation derived for the distribution of the MB statistics based on the assumption of normality will have application in the area of hypothesis testing. Using the derived approximation, it is now possible to analyze samples with respect to normality without needing a table of empirical percentiles (such as Table 2-1) or without having to empirically generate the MB null distribution.

The types of hypothesis tests described in Section 2.2 can now be restated in terms of the approximation of MB. The type of test selected again depends on the information desired about the population from which the sample is taken. Some of these tests will require that a critical region be defined. This is possible when equation (3-2), (with $\gamma = 0$), is solved for MB resulting in the expression

$$MB = \lambda \Big/ \left\{ 1 + \exp \left[-(z - \gamma)/\eta \right] \right\} , \qquad (5\text{-}1)$$

which may then be used to approximate the critical MB values $MBL_{-\alpha}$ and $MBU_{+\alpha}$.

Case 1 - Double Sided MB Test

This test is used to establish the critical region for the acceptance of Ha: a random sample does not come from a normal population. This requires that a critical region be defined based on a desired

confidence level α . The steps required for this test are:

1. Given an error of the first kind; Pr $\{$ MB is in the critical region if Ho is true $\}$ = α ; use equation (5-1) to find the critical region [MBL $_{-\alpha}$, MBU $_{+\alpha}$]. In the equation $\lambda = \ln(n)$ and η and γ are estimated by equation (3-5) and (3-6) for $6 \leq n \leq 50$, or by equations (4-1) and (4-2) for $50 < n \leq 100$. To determine the value of z to be used in equation (5-1), find the area under the standard normal curve where $F(x) = 1 - \alpha/2$ and then find the value of the corresponding standard normal variate z', (which may be found from a standard normal table or by computer subroutine). To find MBL $_{-\alpha}$ use $-z'$ for the value of z in equation (5-1), and for MBU $_{+\alpha}$ use $+z'$ for the value of z.

2. Compute the MB statistic for the random sample of n elements, $x_1, x_2, \ldots, x_n$, using the formula in section 2.1.

3. If the MB value is inside the approximated range [MBL $_{-\alpha}$, MBU $_{+\alpha}$], reject Ha in favor of Ho. If the MB value is in the critical region, then Ho cannot be accepted at this value of α .

For example:

If n = 12 and $\alpha = 0.04$, $\lambda = \ln(n) = 2.485$, $\tilde{\eta} = 2.183$, $\tilde{\gamma} = -2.459$, and z' for $F(x) = 1 - \alpha/2$ is 2.054. MBL $_{-.04}$ = 1.357 from

equation (5-1) with $z = -z'$, and $MBU_{-.04} = 2.206$ from equation (5-1) with $z = +z'$. Therefore, if the MB value calculated for a random sample of size twelve is in the range [1.357, 2.206], Ha may be rejected in favor of Ho at a confidence level of 0.04.

Case 2 - Properties of the Population

Using the approximations to MB it is easy to ascertain "how close" a sample is to being from a normal population and whether the sample exhibits bimodal or long-tailedness (outlier) characteristics. This is dependent upon whether the MB value calculated for the sample is above or below the median value of the null-distribution for the sample size in question, and can be evaluated from the percentile value found using MB approximation.

The steps involved are:

1. Given a random sample of size n with elements $x_1, x_2, \ldots, x_n$, compute the value of the MB statistic, using the formula in section 2.1.
2. Estimate the values of η and γ for the sample size using equations (3-5) and (3-6) if $6 \leq n \leq 50$, or equations (4-1) and (4-2) if $50 < n \leq 100$.
3. Compute the transformed standard normal variate for the sample using equation (3-2).
4. Determine the percentile of the MB statistic by use of a computer subroutine or the standard normal table.

5. Conclusions may be drawn about the sample based on this percentile, p.

   i) If $p > 0.5$, the sample has been drawn from a population exhibiting bimodal characteristics. To establish the confidence level δ_+ for the rejection of Ha: A random sample does not come from a normal population, the following equation is used:

$$\delta_+ = 2(1-p).$$

   ii) If $p < 0.5$, the sample has been drawn from a population exhibiting long-tailed (outlier) characteristics. The confidence level δ_- with which Ha may be rejected for this case is found by the equation:

$$\delta_- = 2p.$$

For example:

   i) Bimodal Population

For n = 12, the value calculated for MB = 1.920. The values η and γ as estimated by equations (3-5) and (3-6) respectively and $\tilde{\eta} = 2.183$ and $\tilde{\gamma} = -2.459$. Then, from equation (3-2), $z = 0.213$ which yields an area under the standard normal curve $p = 0.584$. Since $p > 0.5$, the population from which the sample is drawn has bimodal characteristics. Also, Ha may be rejected in favor of Ho and the sample can be said to come from a normal population with a confidence level of $\delta_+ = 2(1-p) = .832$.

ii) Long-tailed (Outlier) Population

For n = 12 the value calculated for MB = 1.379. The values of η and γ as estimated by equations (3-5) and (3-6) respectively are $\tilde{\eta}$ = 2.183 and $\tilde{\gamma}$ = -2.459. Then, from equation (3-2), z = -1.978 which yields an area under the standard normal curve p = .024. Since p < 0.5, the population from which the sample is drawn has long-tailed (outlier) characteristics. Also, Ha may be rejected in favor of Ho and the sample can be said to come from a normal population at a low confidence level of $\delta_{-} = 2p = .048$.

Case 3 - Single Sided MB Test

The MB approximation can also be used to determine specifically if a sample comes from a bimodal (or long-tailed) population. The alternative hypothesis for this type of test becomes, Ha: A random sample comes from a bimodal (or long-tailed) type of non-normal population. The steps for this test are:

1. Decide on the confidence level, called δ, which is the Pr $\{$ MB is in the critical region if Ho is true $\}$, where $\delta \leq 0.50$.
2. Estimate the values of η and γ for the sample size to be drawn from the population using equations (3-5) and (3-6) if $6 \leq n \leq 50$; or equations (4-1) and (4-2) if $50 < n \leq 100$.
3. Equation (5-1) can then be used to find the critical MB values for the establishment of a critical region. MB

median will always be one of the critical MB values and may be approximated by solving equation (5-1) for MB when z = 0. To find the remaining MB critical value, the value of z in equation (5-1) must be found. This is done by finding the area under the standard normal curve where $F(x) = 1 - \delta$ and then finding the value of the corresponding standard normal variate z'. If a bimodal test is being performed, $z = +z'$ in equation (5-1). If a long-tailed (outlier) test is being performed, $z = -z'$ in equation (5-1). This procedure yields the range [MB median, MB $\underset{-}{+}$ 2 δ].

4. Calculate the value of the MB statistic for the random sample of n elements, $x_1, x_2, \ldots, x_n$, using the formula in section 2.1.

5. If the value of MB calculated in step 4 is inside the range [MB median, MB $\underset{-}{+}$ 2 δ], reject Ha in favor of Ho at a confidence level δ. If the value of MB calculated from the sample is greater (less) than MB $_{+ 2\delta}$ (MB $_{- 2\delta}$), then Ho cannot be accepted at a confidence level δ. The population from which the sample was drawn may then be said to have bimodal (or long-tailed) characteristics at the confidence level δ.

For example:

For n = 12, δ = 0.02, and the alternative hypothesis Ha: A random sample comes from a bimodal type of non-normal population, the values of η and γ as estimated by equations (3-5) and (3-6) are $\tilde{\eta}$ = 2.183 and $\tilde{\gamma}$ = -2.459. Solving equation (5-1) when z = 0, MB median = 1.876. For F(x) = 1- δ = 0.98, z' = 2.054. Since a bimodal type of test is being performed, z = +2.054 in equation (5-1) yielding the value $MB_{+.04}$ = 2.206. The critical region is then outside the range [1.876, 2.206]. If the value of the MB statistic calculated for the sample falls within the range [1.876, 2.206], then Ha may be rejected in favor of Ho. If the MB value is greater than $MB_{+.04}$ = 2.206, then Ho cannot be accepted at a confidence level 0.02 and the population from which the sample was drawn may be said to have bimodal characteristics.

The tests described in this chapter are valid for all sample sizes $6 \leq n \leq 100$ and are dependent only on the size of the sample drawn from the unknown population.

CHAPTER 6

MINIMUM VALUE FOR MB

The minimum value of MB used in developing the approximations to the null and cumulative distributions of MB (see Chapters 3 and 4) was set equal to zero. This is the value obtained for the MB statistic when all the elements, $x_1, x_2, \ldots, x_n$, of the random sample of size n drawn from the unknown population are equal. This type of sample is obviously drawn from a constant population, and is therefore of little value from the point of view of testing for normality. A simple departure from such a sample would occur when all of the elements are equal except one. This type of sample would be expected to yield a MB minimum value other than zero.

The development of MB_{min} for the general case when all elements, $x_1, x_2, \ldots, x_n$, of a random sample of size n are equal except one is as follows:

Let $\quad x_i = x \quad$ for $i = 1, 2, \ldots, (n-1)$

and $\quad x_n = x + n\Delta$.

Then, $\quad \bar{x} = \sum_{i=1}^{n} x_i/n = [(n-1)\ x+x+n\Delta\ ]/n$

$= x + \Delta$.

In order to find y_i where

$$y_i = (x_i - \bar{x})^2 / \sum_{j=1}^{n} (x_j - \bar{x})^2 \quad \text{for } i = 1, 2, \ldots, n,$$

$$(x_i - \bar{x})^2 = [x - (x + \Delta)]^2 = \Delta^2 \quad \text{for } i = 1, 2, \ldots, (n-1),$$

$$(x_n - \bar{x})^2 = [x + n\Delta - (x + \Delta)]^2 = (n-1)^2 \Delta^2 ,$$

and

$$\sum_{j=1}^{n} (x_j - \bar{x})^2 = \sum_{j=1}^{(n-1)} (x_j - \bar{x})^2 + (x_n - \bar{x})^2$$

$$= (n-1)\Delta^2 + (n-1)^2 \Delta^2$$

$$= n(n-1)\Delta^2$$

Substituting these values into the equation for y_i

$$y_i = \Delta^2 / [n(n-1)\Delta^2] = 1/[n(n-1)] \quad \text{for } i = 1, 2, \ldots, (n-1)$$

and $y_n = [(n-1)^2 \Delta^2] / [n(n-1)\Delta^2] = (n-1)/n.$

The minimum value of MB is then

$$MB_{min} = - \sum_{i=1}^{n} y_i \ln y_i$$

which can be rewritten as

$$MB_{min} = - \sum_{i=1}^{(n-1)} y_i \ln y_i - y_n \ln y_n$$

$$= - \sum_{i=1}^{(n-1)} \left\{ \frac{1}{n(n-1)} \ln \left[\frac{1}{n(n-1)} \right] \right\} - \left\{ \frac{n-1}{n} \ln \left[\frac{n-1}{n} \right] \right\}$$

$$= - \frac{1}{n} \ln \left[\frac{1}{n(n-1)} \right] - \frac{n-1}{n} \ln \left[\frac{n-1}{n} \right]$$

which can be simplified to

$$MB_{min} = \ln\left[n/(n-1) \; \frac{n-2}{n} \right] \qquad (6\text{-}1)$$

The MB_{min} values obtained for selected sample sizes from equation (6-1) are tabulated in Table 6-1.

It was shown in reference one, however, that $MB_{min} = \ln(2)$ for samples with symmetrically placed outliers. For $n \leq 6$, this value is less than the value obtained for MB_{min} from equation (6-1).

Therefore,

$$MB_{min} = \begin{cases} \ln(2) & n \leq 6 \\ \ln\left[n/(n-1) \; \frac{n-2}{n} \right] & n \geq 7 \end{cases} \qquad (6\text{-}2)$$

The MB_{min} values defined above were used to develop the Johnson S_B approximation to the null MB distribution to see if a better approximation could be obtained. Equation (3-1) was used with $\epsilon = MB_{min}$ from equation (6-2) and $\epsilon + \lambda = \ln(n)$, and the procedure outlined in Chapter 3 was followed. The smallest Chi-square value obtained for the difference between the Johnson S_B distribution and the empirical MB null distribution in Table 2-1 are exhibited in Table 6-2.

Comparing the Chi-square values in Table 6-2 with the Chi-square values in Tables 3-3 and 4-1, it can be seen that using

$\epsilon = MB_{min}$ (equation 6-2) does not yield as good of an approximation to the MB null distribution for the majority of the sample sizes. There is a significant decrease in the value of Chi-square for n = 4 and n = 5 when $\epsilon = MB_{min}$, but the resulting value of Chi-square for these values of n are still to large to accept the approximations. Therefore, based on these results, it was decided not to pursue any further the use of $\epsilon = MB_{min}$ (equation 6-2) for obtaining an approximation to MB.

TABLE 6-1 MINIMUM VALUES OF MB

$$MB_{min} = \ln\left[n/(n-1)^{\frac{n-2}{n}}\right]$$

n	MB_{min}
4	.837
5	.778
6	.719
7	.666
8	.620
9	.580
10	.545
12	.487
14	.440
16	.403
18	.372
20	.346
25	.295
30	.258
35	.230
40	.208
50	.176
60	.153
70	.135
80	.122
90	.111
100	.102

TABLE 6-2 CHI-SQUARE VALUES FOR GOODNESS OF FIT TEST BETWEEN JOHNSON S_B DISTRIBUTION AND MB EMPIRICAL DISTRIBUTION WHEN $\epsilon = MB_{min}$

n	χ^2
4	956.66
5	179.05
6	271.55
7	125.70
8	105.38
9	79.12
10	42.03
12	9.78
14	12.52
16	14.78
18	25.43
20	19.60
25	13.19
30	15.08
35	10.90
40	10.88
50	9.40
60	12.22
70	12.17
80	4.67
90	8.06
100	13.55

CHAPTER 7
SUMMARY AND CONCLUSIONS

The MB test has been proposed and documented as a test which may be used to determine whether or not a random sample of size n comes from a normal population. The test is easy to apply and is powerful over a wide range of non-normal alternatives. The MB statistic is double sided, and from this property it is also possible to determine if the population being tested for the property of normality has bimodal or long-tailed (outlier) characteristics.

In order to apply the MB test, the distribution of MB statistics under the null hypothesis of normality must be known so that critical regions and confidence levels can be established. Originally this distribution was obtained from samples drawn under the null hypothesis of normality by the method of empirical sampling; a time consuming process. Expressions which can be used to approximate the empirical null distribution and the empirical cumulative distribution of MB have now been derived for samples of sizes six to one hundred. The expressions have been found capable of yielding very close approximations to the MB distributions and are easy to use.

The approximating expressions were derived from the Johnson S_B distribution, a distribution based on the transformation of

a standard normal variate. The Johnson S_B distribution is a four parameter distribution, and equations for three of the parameters were found which are dependent only on the sample size. The fourth parameter, the minimum value of the distribution, was set equal to zero.

The ability to approximate the empirical MB cumulative distribution has valuable application in the area of hypothesis testing. Tests for two hypotheses (1) {Ha: A sample of size n does not come from a normal population} and (2) {Ha: A sample comes from a bimodal (or long-tailed) type of non-normal population} may be considered as alternatives to the null hypothesis {Ho: A sample of size n comes from a normal population}. The Johnson S_B approximation enables the establishment of critical regions and confidence levels for the rejection of the alternative hypothesis chosen in which only the knowledge of the size of the sample drawn from unknown population is required. Furthermore, if the alternative hypothesis is {Ha: A sample comes from a bimodal (or long-tailed) type of non-normal population}, the approximating expressions can be used to establish the confidence with which a population may be considered as having bimodal or long-tailed characteristics.

In this thesis, good approximations were derived from a particular family of distributions - the Johnson family. Although satisfactory results were obtained, a better method of approximating the empirical MB null and cumulative distributions may

exist. A different method may also yield better results for the very small sample sizes and for the non-zero minimum value of MB.

Approximations to the distribution of the MB statistics was not investigated for sample sizes greater than one hundred in this thesis, and in general the behavior of MB for large sample sizes has not been investigated. Also, a comprehensive analysis of the power of the MB test when the approximating expressions were used was not undertaken.

BIBLIOGRAPHY

Theses

1. Berry, G. Lyndon, "A Statistic Based on Standardized Squared Deviations For Testing Normality," Master's Thesis, Lehigh University, 1970.

Reports

2. Bell, Wendell A., Western Electric Company, private communication, with permission.

Articles

3. Johnson, N. L., "Systems of Frequency Curves Generated by Methods of Translation," Biometrika, Vol. 36, 149 (1949), pp. 149-175.

Books

4. Elderton, William P., and Norman L. Johnson, Systems of Frequency Curves, Cambridge University Press, London, England, 1969.

5. Hahn, Gerald J., and Samuel S. Shapiro, Statistical Methods in Engineering, John Wiley and Sons, Inc., New York, 1967.

6. Miller, Irwin, and John E. Freund, Probability and Statistics for Engineers, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1965.

Special Publications

7. International Business Machines, "System/360 Scientific Subroutine Package (360 A-CM-03X) Version III Programmer's Manuel," Publication H20-0205-3, International Business Machines Corporation, 1968.

8. Service Bureau Corporation, "Call/360: STATPACK Statistical Package," The Service Bureau Corporation, New York, 1969.

VITA

PERSONAL HISTORY

Name:	Charlotte A. Klever
Birth Place:	Omaha, Nebraska
Birth Date:	January 27, 1943
Parents:	Mr. and Mrs. Charles F. Klever

EDUCATIONAL BACKGROUND

Peru State College - Education	BS-1964
Lehigh University - Industrial Engineering	MS-1971

ACADEMIC HONORS

Alpha Mu Omega

PROFESSIONAL EXPERIENCE

Western Electric Co., Information Systems Development	1966-1971