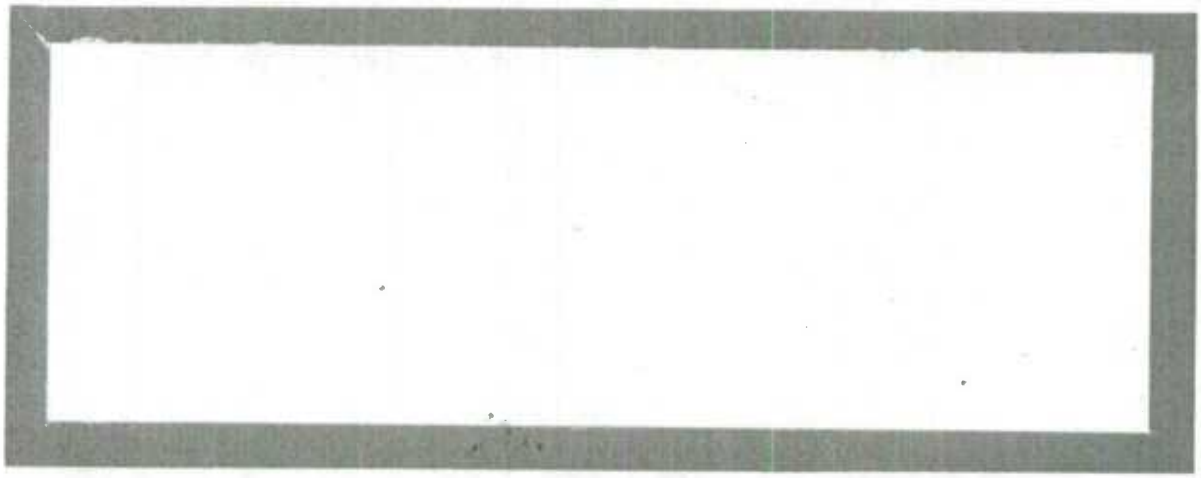
11-617E
no.88-28
c.2

Statistics Canada Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

WORKING PAPER NO. BSMD-88-028E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-88-028E

DIRECTION DE LA MÉTHODOLOGIE

THE USE OF ESTIMATING FUNCTIONS FOR CONFIDENCE INTERVAL CONSTRUCTION: THE CASE OF THE POPULATION MEAN

by

L. Mach

December 1988

STATISTICS CANADA STATISTIQUE CANADA

JAN 9 1995

LIBRARY BIBLIOTHÈQUE

RÉSUMÉ

Dans ce rapport, deux méthodes pour calculer un intervalle de confiance de la moyenne d'un échantillon tiré d'une population normale sont comparées. L'intervalle de confiance A est construit en utilisant la théorie des fonctions d'estimation alors que l'intervalle B est l'intervalle de confiance standard. La probabilité de couverture observée de même que la largeur de ces deux intervalles sont comparées. Les résultats de l'étude indiquent que l'intervalle de confiance A est meilleur que l'intervalle B.

THE USE OF ESTIMATING FUNCTIONS FOR CONFIDENCE INTERVAL CONSTRUCTION: THE CASE OF THE POPULATION MEAN

1. INTRODUCTION

In this report, we describe a study which was conducted to compare two different confidence intervals for a population mean when a simple random sample is selected from a normally distributed population. The theory of estimating functions was used to obtain the first interval. The second interval is the usual confidence interval based on the approximately normal distribution of the sample mean.

The purpose of the study was to demonstrate how the theory of estimating functions works in the case where a well established technique already exists. Proof that the estimating function theory works well in interval estimation for the mean in this instance should stimulate interest in conducting empirical studies to test the performance of the theory in cases where new interval estimation methods are needed (eg. estimation of percentiles, ratios...).

The theory of estimating functions is briefly summarized in Chapter 2. The two methods of constructing confidence intervals for the mean are described in Chapter 3. The length and the coverage of the two intervals are compared in Chapter 4. Chapter 5 summarizes the results of the study.

2. THEORY OF ESTIMATING FUNCTIONS

The theory of estimating functions is discussed by Godambe and by Godambe and Thompson ([3], [4], [5], [6], [7]). Let $S = y_1, y_2, \ldots y_n$ denote the observed values of a simple random sample $Y_1, Y_2, \ldots Y_n$ selected from a population with an unknown parameter θ. Then a real function $g(S,\theta)$ is called an estimating function if an estimate of the parameter θ can be obtained by solving the equation

$$g(S,\theta) = 0 \qquad (2.1)$$

for θ. The equation (2.1) is called an estimating equation.

An estimating function g is said to be unbiased if

$$E[\, g(S,\theta) \,] = 0 \qquad (2.2)$$

for all permissible values of θ, where E denotes the expectation. In [3], a regular estimating function and an optimum estimating function are defined in full. Briefly, a regular estimating function is an estimating function which satistfies four general conditions given in [3]. An optimum estimating function $g^*(S,\theta)$ is a regular estimating function for which

$$E(g^2|\theta)/[E\{(\partial g/\partial\theta)|\theta\}]^2 \qquad (2.3)$$

is minimized among all regular estimating functions and for all permissible values of θ. Hence, by this definition, an optimum estimating function is the estimating function which has a small variance and, at the same time, $E[\, g(S,\theta+\Delta\theta)|\theta\,]$ is as far away from 0 as possible (ie. g which is sensitive to the changes of θ).Under the usual regularity conditions it is proved in [3] that the maximum likelihood estimation is optimum if only one population parameter is unknown.

If S is selected from a population with two or more unknown parameters and only one of them is to be estimated, the maximum likelihood fails to be the optimum. The optimum estimating function to estimate θ_1, when S is selected from a population with two unknown parameters, θ_1 and θ_2, is given in [4]. It is a linear combination of derivatives of log likelihood function and constants depending on θ_1 and θ_2 (such that the resulting g^* is independent of θ_2). For instance, let S be selected from a normal population with an unknown mean μ and an unknown variance σ^2. Then solving the estimating equation $g^*(S,\sigma^2)=0$ gives the optimal estimate of population variance, $\Sigma(y_i-\bar{y})^2/(n-1)$, while the maximum likelihood estimate, $\Sigma(y_i-\bar{y})^2/n$ is biased.

A linear estimating function is defined in [7] as

$$g(S,\theta) = \sum_{i=1}^{n} \{\phi_i(y_i,\theta)\}a_i(\theta), \qquad (2.4)$$

where ϕ_i is a real function with $E(\phi_i) = 0$, and $a_i(\theta)$ is any real function of θ, i=1,...,n. A linear estimating function is said to be linearly optimal if (2.3) is minimized for all linear estimating functions. As shown in [7],

$$g(S,\theta) = \sum_{i=1}^{n} \{\phi_i(y_i,\theta)\} \qquad (2.5)$$

is linearly optimal if

$$E(\partial\phi_i/\partial\theta)/E(\phi_i^2) = C, \qquad (2.6)$$

where C is a constant. (Often, the optimal estimating function is in the form of (2.5). However, in some cases, although no optimal estimating function exists, a linear optimal estimating function may exist.) Also, if $\phi_i=\phi$ and (2.6) holds, then (2.5) is optimal.

The theory of estimating functions can be extended to estimation of a survey population parameter when more complex sampling designs are used and to estimation of several population parameters. This is also discussed by Godambe and Thompson in [7]. Some applications of the theory were discussed by other authors as well. For instance, Fieller [2] used the approach to derive confidence limits for a ratio, and Binder [1] applied the theory for estimating the variance of estimated parameters based on complex sample designs from finite populations.

3. TWO METHODS OF CONSTRUCTING CONFIDENCE INTERVALS FOR THE MEAN

3.1 Confidence Interval for the Mean Based on the Theory of Estimating Functions

We consider a specific example of an estimating function. A simple random sample $S = y_1, y_2, \ldots y_n$ is selected from an infinite population with mean μ. The optimum estimating function to estimate μ is

$$g(S,\mu) = \sum_{i=1}^{n} (y_i - \mu). \tag{3.1}$$

The estimating equation $g(S,\mu) = 0$ solved for μ yields the usual sample mean as an estimator of μ.

Let $v\{g(S,\mu)\}$ be a consistent estimator of the variance of the estimating function $g(S,\mu)$. Then, it implies from some basic limit theorems (see [9], for instance), that

$$g(S,\mu)/(v\{g(S,\mu)\})^{1/2} \tag{3.2}$$

approaches the standard normal variable $N(0,1)$ as n increases, where μ is the true mean of the population.

We use $\sum_{i=1}^{n} (y_i - \mu)^2$ as a consistent estimator of the variance of $g(S,\mu)$. Hence

$$\sum_{i=1}^{n} (y_i - \mu) / \left(\sum_{i=1}^{n} (y_i - \mu)^2 \right)^{1/2} \tag{3.3}$$

approaches $N(0,1)$ as n increases.

Let $z_{\alpha/2}$ denote the $(1 - \alpha/2)$th percentile of the standard normal variable. Then

$$\Pr(-z_{\alpha/2} < \Sigma(y_i-\mu)/[\Sigma(y_i-\mu)^2]^{1/2} < z_{\alpha/2})$$

$$= \Pr(-z_{\alpha/2} < n(\bar{y}-\mu)/[(n-1)s^2+n(\bar{y}-\mu)^2]^{1/2} < z_{\alpha/2}) = 1 - \alpha, \tag{3.4}$$

where $\bar{y}$ and s^2 denote the sample mean and variance respectively.

After some algebraic operations, the inequality in parenthesis in (3.4) can be rewritten as

$$n(\bar{y}-\mu)^2/s^2 < z_{\alpha/2}^2[(n-1)/(n-z_{\alpha/2}^2)], \qquad \text{if } (n-z_{\alpha/2}^2) > 0.$$

Hence if $n > z_{\alpha/2}^2$ the left side of (3.4) can be expressed as

$$\Pr\{-z_{\alpha/2}\sqrt{(n-1)/(n-z_{\alpha/2}^2)} < \sqrt{n}(\bar{y}-\mu)/s < z_{\alpha/2}\sqrt{(n-1)/(n-z_{\alpha/2}^2)}\}. \tag{3.5}$$

If the variable to be measured, Y, possesses the normal probability distribution, then the middle expression in (3.5) has the Student's t

distribution with (n-1) degrees of freedom. Thus we can obtain the exact probability $(1 - \tau)$, $0 \leq \tau \leq 1$:

$$\Pr\{-z_{\alpha/2}\sqrt{(n-1)/(n-z_{\alpha/2}^2)} < t_{n-1} < z_{\alpha/2}\sqrt{(n-1)/(n-z_{\alpha/2}^2)}\} = 1 - \tau. \quad (3.6)$$

By using the pivotal method for constructing confidence intervals we obtain the claimed $(1 - \alpha)100$ % confidence interval for the mean μ

$$\bar{y} \pm z_{\alpha/2}\sqrt{(n-1)/(n-z_{\alpha/2}^2)}\, s/\sqrt{n}. \quad (3.7)$$

We will call the interval given in (3.7) interval A. When the population is normally distributed then the true coverage probability of interval A is $(1 - \tau)$.

The interval (3.7) is the $(1-\alpha)100\%$ confidence interval for the mean μ based on the theory of estimating functions when the true parameter μ was used in the estimator of the variance of $g(S,\mu)$. It should be noted that a different interval for μ based on the theory of estimating functions can be obtained by choosing a different consistent estimator of the variance of $g(S,\mu)$, $v\{g(S,\mu)\}$.

3.2 Classical Confidence Interval for the Mean

We compare the interval (3.7) with the classical confidence interval for the mean. If the sample is large (usually $n > 30$), then the classical $(1-\alpha)100\%$ confidence interval for the mean μ is constructed by using the values of the standard normal variable instead of the t-values:

$$\bar{y} \pm z_{\alpha/2}s/\sqrt{n}. \quad (3.8)$$

We will denote the interval given in (3.8) as interval B. Again, if the sample is selected from a normal population, then the true coverage probability of interval B is $(1 - \delta)$, $0 \leq \delta \leq 1$:

$$\Pr\{-z_{\alpha/2} < \sqrt{n}(\bar{y} - \mu)/s < z_{\alpha/2}\} = \Pr\{-z_{\alpha/2} < t_{n-1} < z_{\alpha/2}\} = 1 - \delta. \quad (3.9)$$

4. COMPARISON OF THE TWO CONFIDENCE INTERVALS FOR THE MEAN

4.1 Comparison of the Length of the Intervals A and B

The length of the confidence intervals A and B is compared. Let L(.) denote the length of the confidence interval. Then

$$L(A) = (2z_{\alpha/2}s/\sqrt{n})[\sqrt{(n-1)/(n-z_{\alpha/2}^2)}], \quad (5.1)$$

$$L(B) = 2z_{\alpha/2}s/\sqrt{n}. \quad (5.2)$$

For $z_{\alpha/2}^2 > 1$ (ie. for $1 - \alpha > .6826$) the following relationship holds:

$$L(B) < L(A). \qquad (5.3)$$

Hence, the confidence interval A is wider than the classical confidence interval B in most practical situations. The difference between the length of A and B increases as the confidence level increases and decreases as n becomes large. However, it is known (and will also be discussed in 4.2), that interval B is not wide enough. Its true coverage probability is always less than its stated confidence level $(1 - \alpha)$. Thus, an improved symmetric confidence interval must be wider than interval B.

4.2 Comparison of the True Coverage for the Intervals A and B

The coverage probabilities of the two confidence intervals for the population mean were also compared for simple random samples from normal populations. To this end, the true coverage probabilities $(1-\tau)$ and $(1-\delta)$ were computed for the intervals A and B, respectively, for different claimed confidence levels $(1-\alpha)$ and different sample sizes n.

Graphs 1.1, 1.2 and 1.3 show the true coverage probabilities of intervals A and B against sample size n for a claimed confidence level of .90, .95 and .99, respectively. The coverage probability for interval A is slightly below the preassigned confidence level on Graph 1.1 and slightly above the preassigned confidence level on Graphs 1.2 and 1.3. On the other hand, the coverage probability for interval B is always below the stated confidence level. The difference between the true and the claimed coverage probabilities decreases with increasing n . This difference is considerably smaller for interval A than for interval B, especially for $(1-\alpha) = .90$ and .95.

Graphs 2.1, 2.2 and 2.3 show the true coverage probabilities of intervals A and B against claimed confidence level for a sample of size 15, 30 and 50, respectively. The true coverage probability for interval A is below the claimed confidence level for smaller values of $(1-\alpha)$ and above the claimed confidence level for $(1-\alpha) > .90$. The true coverage probability for interval B is always smaller than the claimed confidence level and the coverage probability for A. As n increases the distance between the true and the claimed coverages decreases for both intervals, but the true coverage for interval A is always closer to the claimed confidence level than that of interval B.

5. CONCLUSION

Two methods of constructing confidence interval for the population mean are described. These confidence intervals are compared when a simple random sample of units is selected from a normally distributed population. Interval A is constructed by using the theory of estimating functions.

Interval B is the usual confidence interval based on the approximately normal distribution of the sample mean. Since both intervals are symmetric, centred at the sample mean, the criterion used to compare the performance of the two intervals is the coverage probability only. The coverage probability of A is always closer to the claimed confidence level than the coverage probability of B. For $(1-\alpha)>.90$, A provides slightly conservative intervals. Otherwise, the true coverage is always below the stated confidence level. Hence we can conclude that interval A is a better confidence interval for a normal population mean than interval B. It should be noted that interval B is usually used for large samples only ($n>30$). The Student's t distribution is used to construct confidence intervals when small samples are selected.

The major objective of the study was to investigate how the theory of estimating functions works in the case where a well established technique already exists, rather than to select the better method of the two. The study shows that the confidence intervals based on the estimating functions theory provide good interval estimators of a mean of a normally distributed population. This result encourages us to investigate the performance of confidence intervals based on the theory of estimating functions in cases where new interval estimation methods are needed (eg. estimation of percentiles, ratios etc.).

ACKNOWLEDGEMENT

I would like to thank Dr. David A. Binder for suggesting the study and for his guidance during the preparation of this paper. I am also grateful to John Kovar for his valuable advice, suggestions and comments. My thanks go to Nancy Darcovich for proofreading the manuscript and help with programming.

REFERENCES

[1] Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. **International Statistical Review**, 51, 279-292.

[2] Fieller, E. C. (1932). The distribution of the index in a normal bivariate population. **Biometrika**, 24, 428-440.

[3] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. **Annals of Mathematical Statistics**,31, 1208-1211.

[4] Godambe, V. P., and Thompson, M. E. (1974). Estimating equation in the presence of a nuisance parameter. **The Annals of Statistics**, 2, 568-571.

[5] Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. **Biometrika**, 63, 277-284.

[6] Godambe, V. P., and Thompson, M. E. (1978). Some aspects of the theory of estimating equations. **Journal of Statistical Planning and Inference**, 2, 95-104.

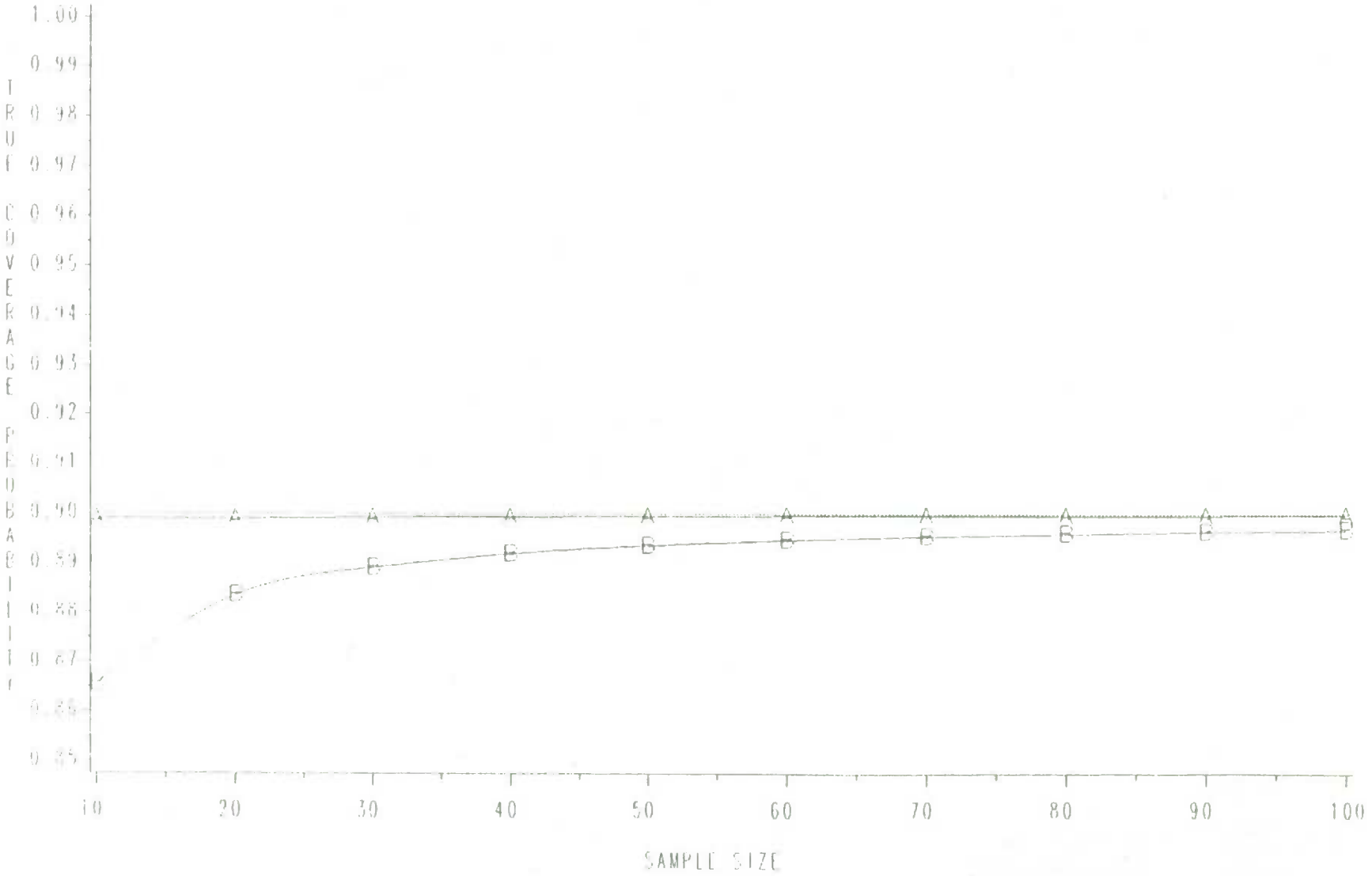
[7] Godambe, V. P., and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. **International Statistical Review**, 54, 127-138.

[8] Rao, C. R. (1973). **Linear Statistical Inference and its Applications**, 2nd edition, New York: Wiley.

[9] Roussas, G. G. (1973). **A First Course in Mathematical Statistics**, Addison - Wesley Publishing Company, Inc.

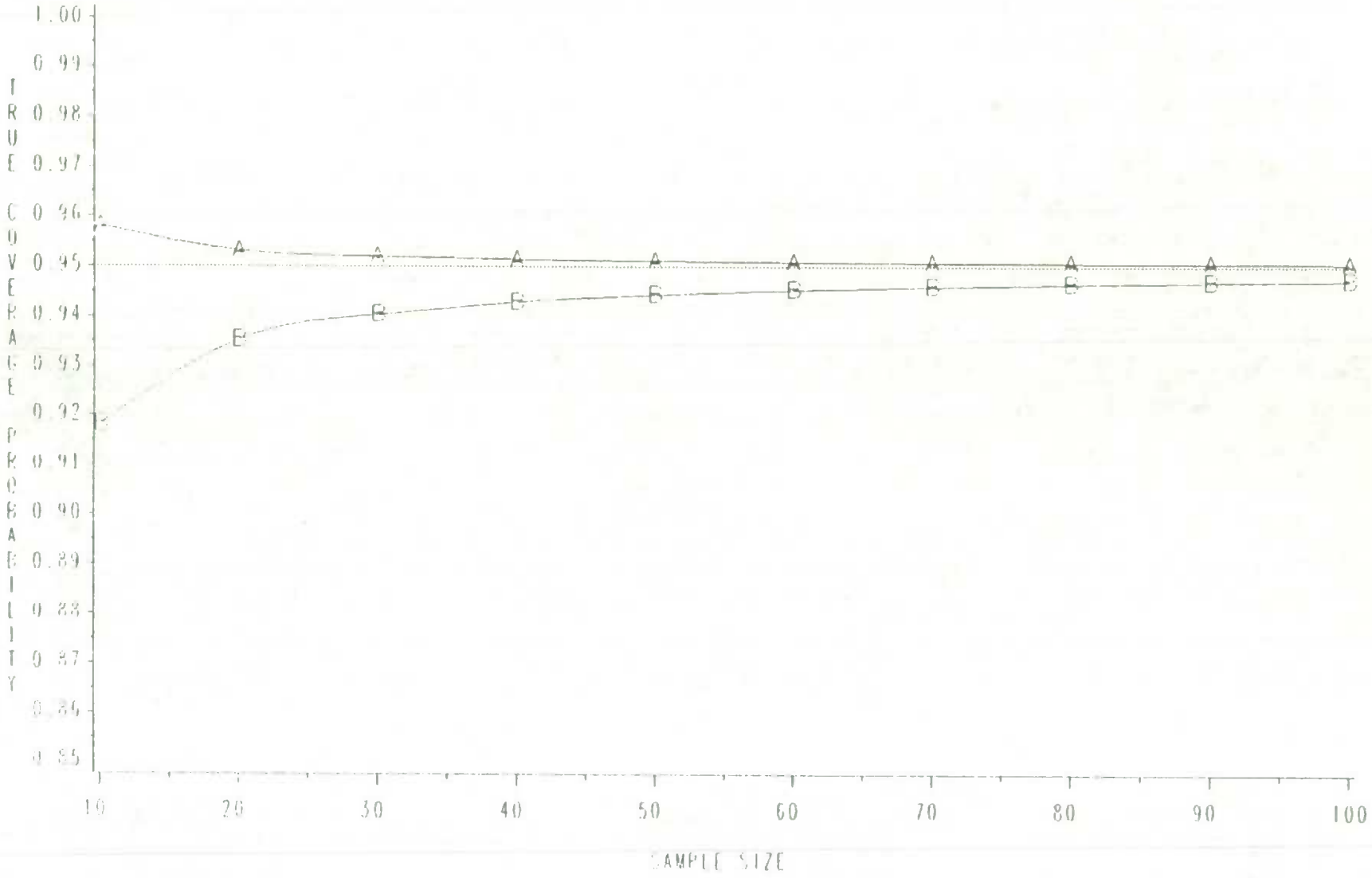
GRAPH 1.1

COMPARISON OF TRUE COVERAGE FOR INTERVALS A AND B
STATED CONFIDENCE LEVEL IS 90 %



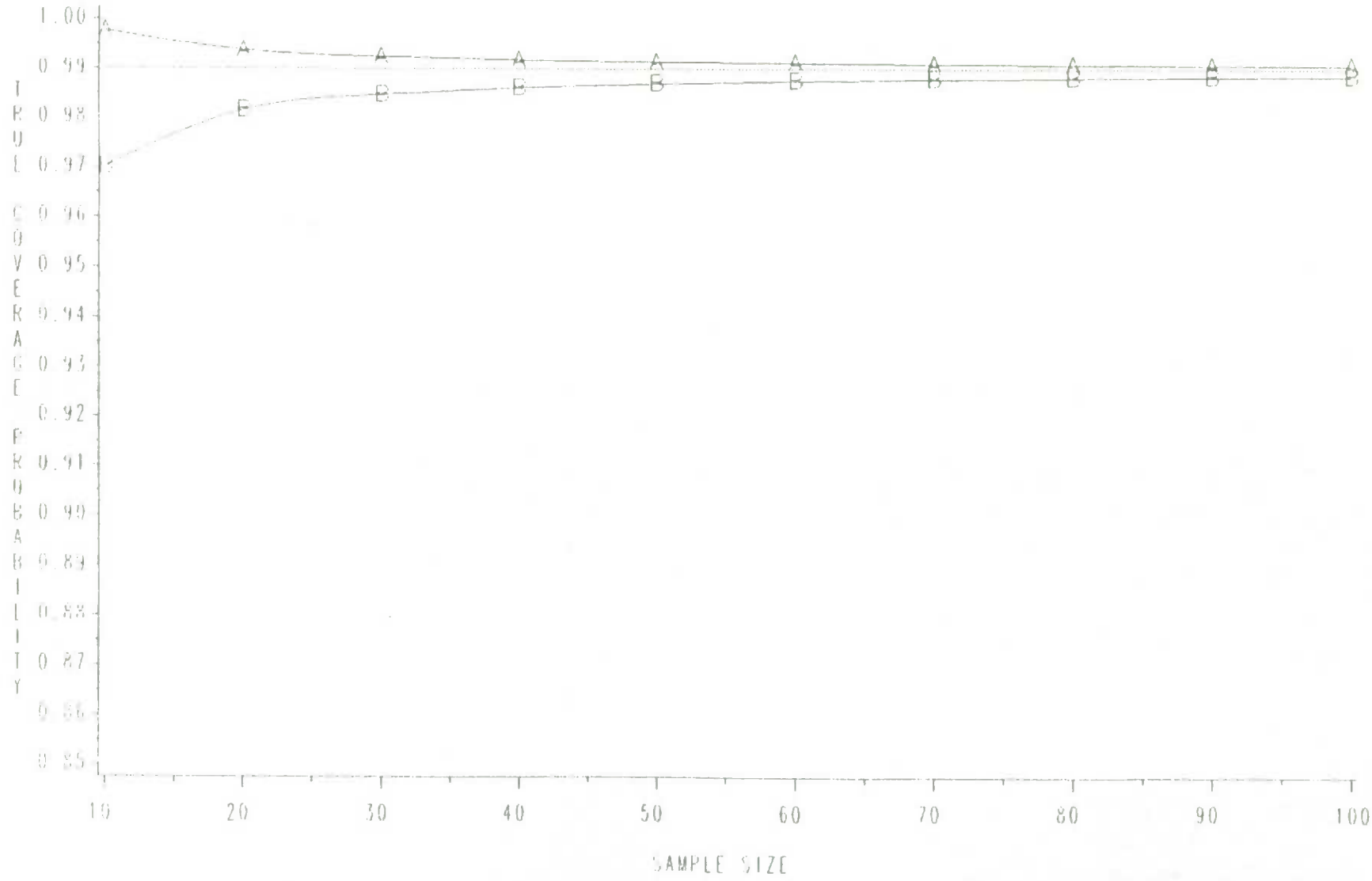
GRAPH 1.2

COMPARISON OF TRUE COVERAGE FOR INTERVALS A AND B
STATED CONFIDENCE LEVEL IS 95 %



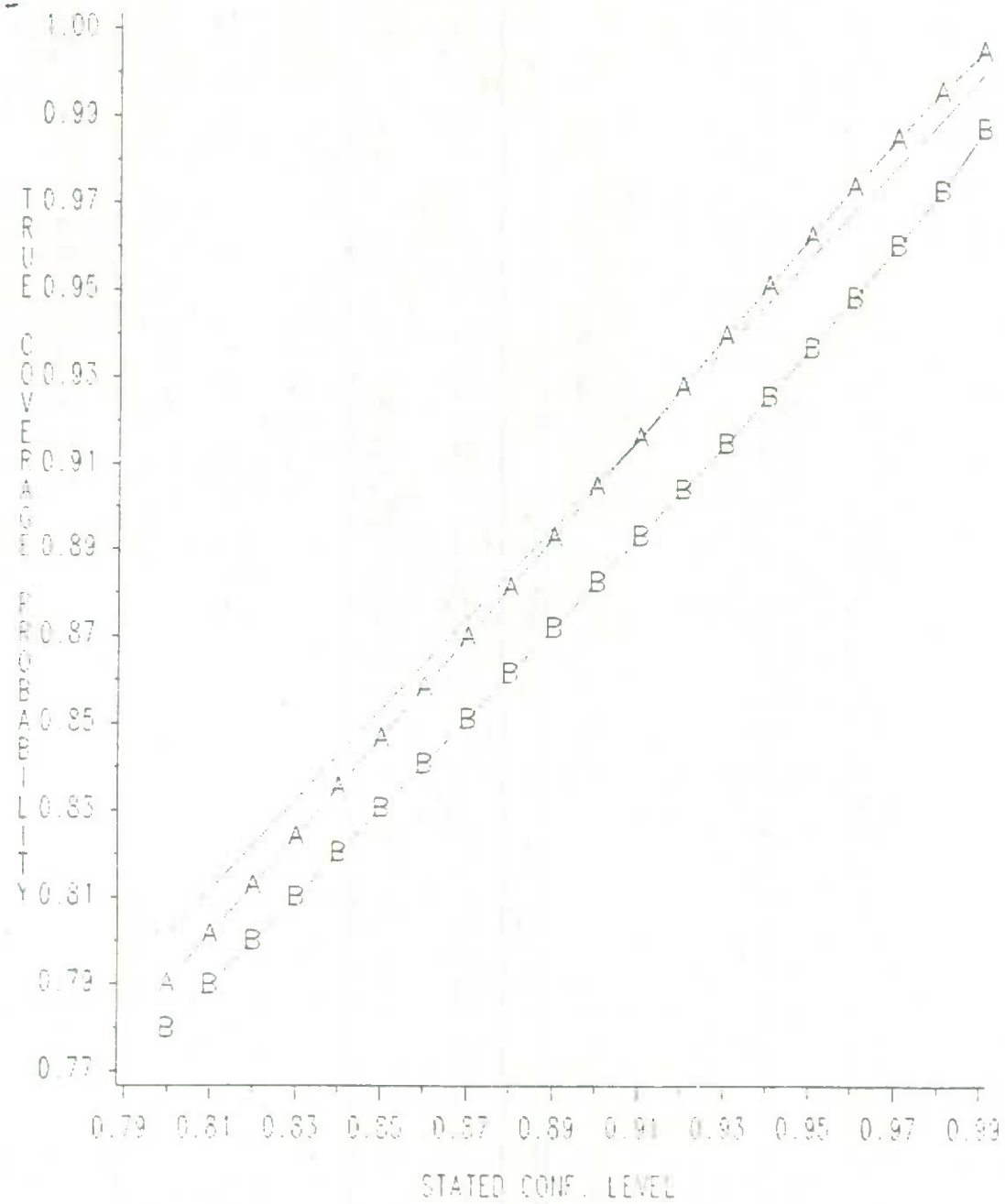
GRAPH 1.3

COMPARISON OF TRUE COVERAGE FOR INTERVALS A AND B
STATED CONFIDENCE LEVEL IS 99 %



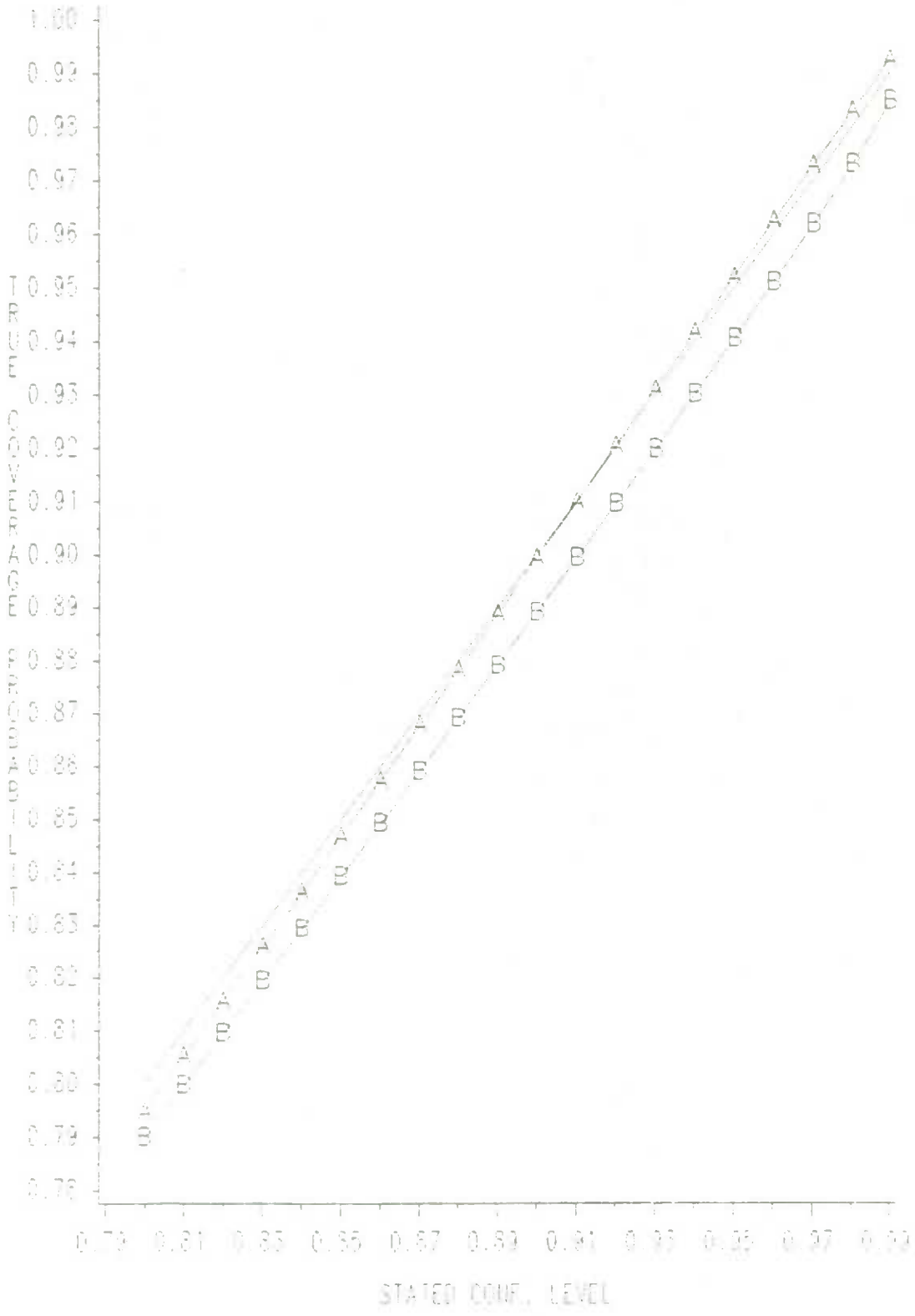
GRAPH 2.1

COMPARISON OF TRUE COVERAGE FOR INTERVALS A AND B
SAMPLE SIZE IS 15



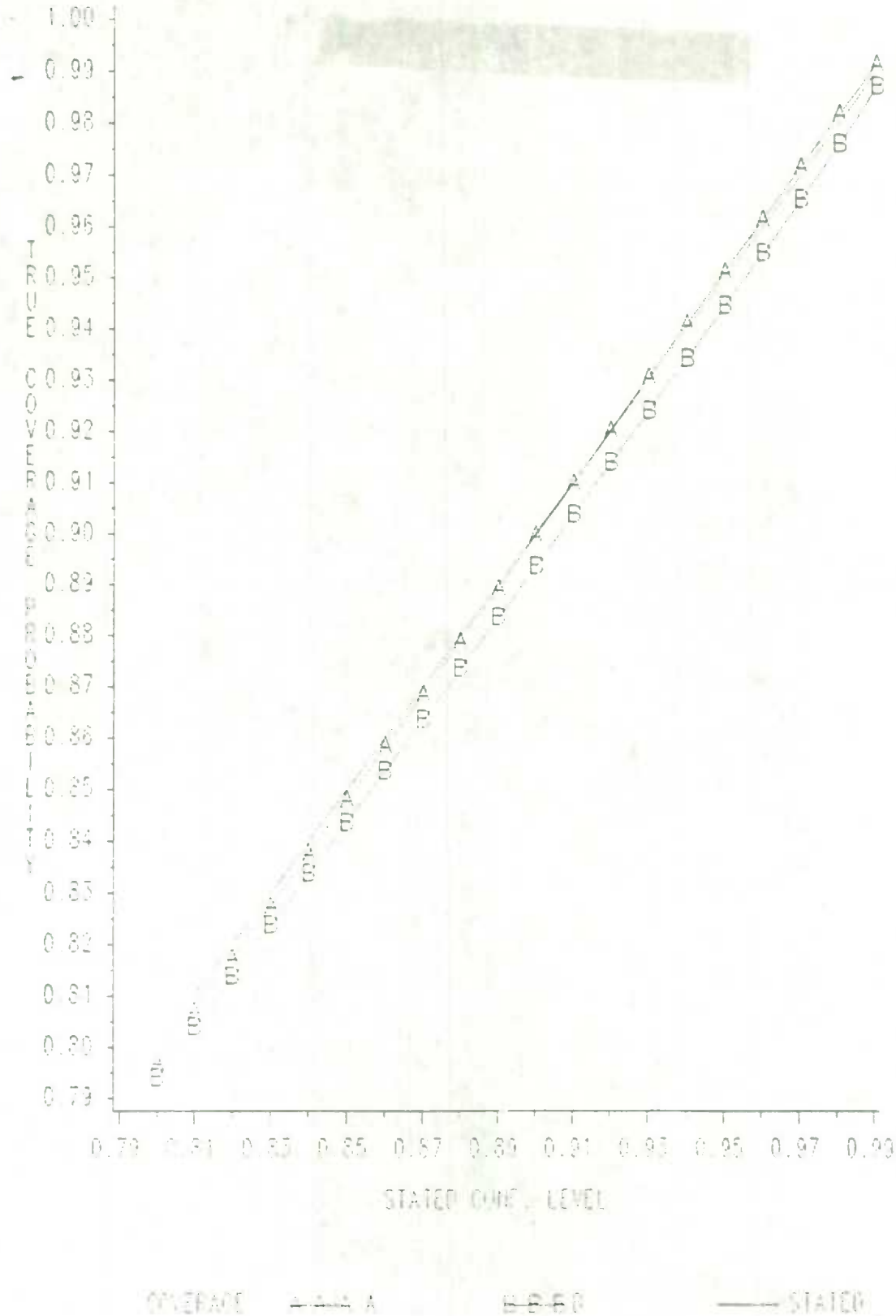
GRAPH 2.2

COMPARISON OF TRUE COVERAGE FOR INTERVALS A AND B
SAMPLE SIZE IS 30



GRAPH 2.3

COMPARISON OF TRUE COVERAGE FOR INTERVALS A AND B
SAMPLE SIZE IS 50



STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010180968

OOS