LEVEL OF SIGNIFICANCE AND POWER OF TWO COMMONLY USED PROCEDURES FOR COMPARING MEAN VALUES BASED ON CONFIDENCE INTERVALS

NANCY C. H. LO
Southwest Fisheries Science Center
National Marine Fisheries Service, NOAA
P.O. Box 271
La Jolla, California 92038

ABSTRACT

Confidence intervals (CIs) are frequently used to compare true means of two populations in the following ways: (A) If two 95% CIs are overlapping, then it can be concluded that the two population means are the same. (B) When only one CI is available, it can be concluded that two means are equal if one sample mean is within the 95% CI of the other mean. But the level of significance (α) of these two procedures does not always equal the intended 5%. The statistical power of these two procedures is unknown. This paper recommends another statistical procedure: (C), which is based on the CI of the difference (d) of two population means: CI(d). In this simulation study, the actual level of significance and the statistical power of these three procedures are computed for equal sample sizes. Statistical distributions considered are normal, Poisson, gamma, and lognormal. The simulation results indicate that the α value is 0.005 averaged over three continuous distributions (for Poisson, it is 0.06) for procedure A; 0.17 for procedure B; and 0.05 for CI(d). Thus, when the true means are indeed different, B is the most powerful procedure, and A is the least powerful procedure.

RESUMEN

Los intervalos de confianza (IC) para comparar la media verdadera de dos poblaciones se usan frecuentemente de dos maneras. (A) Si dos IC al 95% traslapan, se puede concluir que la media de las dos poblaciones no es la misma. (B) Cuando sólo se dispone de un IC, puede concluirse que las dos medias son iguales si una de éstas se encuentra dentro del IC de 95% de la otra media. Sin embargo, el nivel de significancia (α) de estos dos procedimientos no siempre es igual al 5% deseado. La potencia estadística de estos dos procedimientos es desconocida. Esta contribución recomienda otro procedimiento estadístico: (C), basado en los IC de las diferencias (d) de dos medias poblacionales: CI(d). En este estudio, usamos simulaciones y calculamos el nivel de significancia real y la potencia estadística de estos tres procedimientos (para muestras de tamaño igual). Se usaron distribuciones normal, Poisson, gama y lognormal. Los resultados de las simulaciones indicaron que para el procedimiento A los valores α promediados en tres distribuciones contínuas fué de 0.005 (0.06 para la distribución Poisson), 0.17 para el procedimiento B y 0.05 para el procedimiento C. Consecuentemente, cuando las medias verdaderas son diferentes, el procedimiento con mayor potencia es el B, mientras que el de menor potencia es el A.

[Manuscript received February 7, 1994.]

INTRODUCTION

In the scientific literature, summary statistics such as averages and standard errors are often used to construct confidence intervals (CIs) for the true mean under the assumption of the normal distribution of the sample mean. Frequently, CIs are used to compare means of two populations in the following ways; (A) If two 95% CIs are overlapping, it can be concluded that the two population means are the same. (B) When only one CI is available, it can be concluded that two means are equal if one sample mean is within the 95% CI of the other mean. Although these two procedures are convenient and popular ways of making inferences about population means, their level of significance (α) does not always equal the intended 5%.

The correct statistical procedures should be those based on the difference of two sample means: for example, t statistics and the CI for the difference of population means: CI(d). When CI(d) is used, and if the CI(d) contains zero, it can be concluded that the two population means are the same. Thus there are actually three procedures to be considered: procedure A is the overlapping of two CIs; procedure B is the inclusion of one sample mean in the CI from the second sample; and procedure C is the CI(d) based on the difference of two sample means assuming normal distribution.

The procedure used is important because it affects conclusions. Take, for example, Hunter and Leong's (1981) comparison of the mean batch fecundity of northern anchovy (15–19g) that matured in the laboratory and in the sea (figure 1):

Locality	n	Mean	(2SE)	95% CI
Laboratory	38	8910	(1210)	(7700, 10120)
Sea	17	6800	(1150)	(5650, 7950)
Difference		2110	(1669)	(441, 3779)

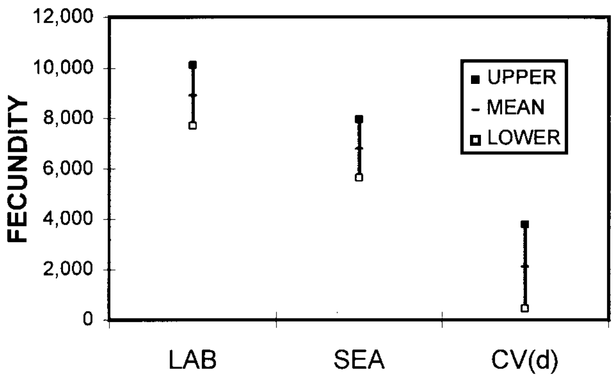

Figure 1. Confidence intervals based on the individual sample mean fecundity of anchovy that matured in the laboratory and in the sea. The CI(*d*) is the confidence interval based on the difference of two sample mean fecundities. Low and high are the lower and upper 95% confidence limits.

The two mean batch fecundities are not significantly different if procedure A is used, but they are significantly different if B or C is used. Clearly, procedure C is the preferred choice, since it is known that the confidence level is 95%. If A or B must be used, however, the investigator should be aware that the level of significance can affect the result.

The purpose of this paper is to compare these three procedures. A simulation study was conducted to compute the actual confidence level $(1-\alpha)$, and the power of procedures A and B was compared to that of procedure C. Note that the level of significance (α) is the probability of claiming that the two population means are different by rejecting the null hypothesis that two population means are equal when, in fact, they are the same. This is a wrong conclusion and is the so-called type I error. The power is also the probability of rejecting the null hypothesis that two means are equal, when the two population means are indeed different. Thus the power is the probability of making the right decision at the expense of α value. While the α value is normally predetermined, the power of a procedure depends not only on the α value, but also on the magnitude of the difference between two population means and the sample size. Statistical power has been recognized as an important element in evaluating fishery population estimation procedures (Peterman 1990; Solow and Steel 1990). But the power must be considered together with the level of significance.

In the simulation, all 95% CIs were computed on the assumption of normally distributed sample means for four underlying distributions: normal, Poisson, gamma, and lognormal. The density functions are given in the appendix. Different statistical distributions for various sample sizes ranging from 5 to 100 were included in the simulation to check the robustness of the CI(*d*). The CI(*d*) is based on the normal assumption, which is valid only for large sample size. Here, CI(*d*) is robust if it maintains a 95% confidence level as intended.

In this paper I do not provide the optimal CI of the difference of population means for each distribution. I refer the reader to Barr 1969 and Nelson 1989 for normal; Casella and Robert 1989 for Poisson; Withers 1991 for gamma; Land 1988 for lognormal; Douglas 1993 and Weerahandi 1993 for a generalized CI; and Beal 1989 for CI and sample size in general. Barr (1969) and Nelson (1989) dealt with the overlapping problems for the normal case only. Ideally, distribution-specific CIs should be sought, but some of the procedures are complex and difficult to apply. Although the normal-based CIs are convenient, their limitation should be recognized. They should be used with caution, particularly when sample size is small.

METHODS

Suppose that one sample of size (n_i) is taken from each of two populations with mean (μ_i) and standard deviation (σ_i), $i = 1, 2$. The goal is to determine whether the population means are equal, i.e., $\mu_1 = \mu_2$. Each of the three procedures would lead to the conclusion that the two averages are not significantly different if:

A. CI_1 and CI_2 are overlapping, where $CI_i =$

$$\bar{x}_i \pm t_{df_i \alpha/2} s_{\bar{x}_i} \quad for\ i = 1, 2 \tag{1}$$

where df_i (degree of freedom) $= n_i - 1$, and $s_{\bar{x}_i}$ is the standard error of the sample mean.

B. The sample mean from one data set is within the CI computed from the other data set (CI_2).

$$\bar{x}_1 \in CI_2 = \bar{x}_2 \pm t_{df_2,\alpha/2} s_{\bar{x}_2} \tag{2}$$

C. The CI(*d*) of the difference of the population means contains zero. The CI(*d*) was computed from the difference of two sample means and the standard error of the difference. This is the confidence interval based on the normal distribution of the sample mean. The 95% CI for $d = \mu_1 - \mu_2$ is

$$CI_d = (\bar{x}_1 - \bar{x}_2) \pm t_{df,\alpha/2} s_{\bar{x}_1 - \bar{x}_2} \tag{3}$$

where $df = (n_1 - 1 + n_2 - 1)$ if variances are equal. If variances are not equal, the formula for the d.f. is available from statistics books (Zar 1984).

The formulas related to the standard error of the sample mean have not been given because they can be found in any statistics reference. Equations 1 to 3 are used for three underlying distributions. In all cases, sample sizes were set to be equal, i.e., $n_1 = n_2$.

In the simulation, data are generated from each of the distribution functions for sample sizes ranging from 5 to 100 (actual sample size may differ for different distri-

TABLE 1
Mean (μ) and Standard Deviation (σ) for Each Distribution Used in the Simulation

Distribution	Parameters*
Normal	μ_1 = 1, and μ_2 = 1, 1.5, 2, 3, 4, and 5 $\sigma_1 = \sigma_2 = 2$
Poisson	mean = variance: μ_1 = 0.5, and μ_2 = 0.5, 1, 2, 5, and 10
Gamma	b(scale parameter) = 1 $\mu_1 = bc = 3$ and μ_2 = 0.5, 1, 1.5, 2, 2.5, and 3 where c is the shape parameter
Lognormal	μ_1 = 0 and μ_2 = 0, 0.5, 1, and 2 $\sigma_1 = \sigma_2 = 1$

*μ_1 is the mean under the H_0 and μ_2 is the mean under the H_a.

butions). To compute the α values, two independent samples were generated from distributions with same mean value. To compute β values, two independent samples were generated from two distributions with different mean values. One thousand iterations (m) were run for each comparison, and the actual level of significance (α) under the null hypothesis H_0: $\mu_1 = \mu_2 = \mu$ was computed for various μ values as:

$$\alpha_A = (\text{number of two CIs not overlapping})/m \quad (4)$$

$$\alpha_B = (\text{number of } \bar{x}_1 s \text{ is not contained in } CI_2)/m \quad (5)$$

and

$$\alpha_C = (\text{number of CI}(d)\text{s not containing zero})/m \quad (6)$$

The power of the three procedures was also computed for various sample sizes under the alternative (H_a): $\mu_2 \neq \mu_1$, where μ_1 was kept constant and μ_2 varied. The confidence level for each individual CI was 95%. Under H_a: $\mu_2 \neq \mu_1$, data were generated from two populations, one with mean μ_2, and one with μ_1. I computed β_A, β_B, and β_C in the same way as the α values (equations 4–6).

The parameters for each distribution are given in table 1. The α values were computed from two samples, each taken from populations with identical mean values indicated by μ_2 in table 1. Samples from the population with mean values equal to μ_2 are compared with samples from the population with mean values equal to μ_1 to compute the power.

RESULTS

Normal Distribution

The α values for $\mu_1 = \mu_2$ = 1, 1.5, 2, 3, 4, and 5 were computed for sample sizes 5 to 100, even though the difference of μ_1 and μ_2, not the actual values of μ_1 and μ_2, is relevant. The standard deviation is set at 2 (table 2). For procedure A, the computed α values are less than 0.05, with an average of 0.005. For procedure B, the opposite is true, and the computed α values are greater than 0.05, with an average of 0.15. For procedure C, the α values are close to 0.05, as expected. In procedure A, the true α values are not affected by the sample sizes, as they are in procedure C. In procedure B, however, the level of significance is affected by the sample size. The α values for sample sizes ≤10 ranging from 0.12 to 0.14 are smaller than those ranging from 0.14 to 0.17 for larger sample sizes.

TABLE 2
Level of Significance (α), Confidence Level ($1-\alpha$), and Overall Power for Three Procedures, for Each Statistical Distribution

	Procedures			
	A	B	C	Hypothesis
Normal				
α	0.005	0.15	0.05	H_0: $\mu_1 = \mu_2$
$1-\alpha$	0.995	0.85	0.95	
Power relative to option C	low	high		H_a: $\mu_2 \neq \mu_1$
Poisson				
α	0.06	0.17	0.05	H_0: $\mu_1 = \mu_2$
$1-\alpha$	0.94	0.83	0.95	
Power relative to option C	low	high		H_a: $\mu_2 \neq \mu_1$
Gamma				
α	0.006	0.16	0.05	H_0: $\mu_1 = \mu_2$
$1-\alpha$	0.994	0.84	0.95	
Power relative to option C	low	high		H_a: $\mu_2 \neq \mu_1$
Lognormal				
α	0.004	0.19	0.05	H_0: $\mu_1 = \mu_2$
$1-\alpha$	0.996	0.81	0.95	
Power relative to option C	low	high for n < 30		H_a: $\mu_2 \neq \mu_1$

The power values were computed for μ_2 = 1, 1.5, 2, 3, 4, and 5, compared to μ_1 = 1. Again, σ = 2 and sample size = 10, 20, . . . , 50, 100. Procedure B is the most powerful of the three, and A is the least powerful, regardless of sample sizes, if the population means are different (table 2 and figure 2). For example, when μ_2 = 2, compared to μ_1 = 1, for sample size 30, the power is 0.18 for procedure A, 0.69 for procedure B, and 0.5 for procedure C. The difference in power among the three distributions is more pronounced for a small sample size and a small difference in mean values.

Poisson Distribution

The Poisson distribution (equation A2) is often used to model the distribution of counts of rare events that are randomly distributed in time and space, e.g., the number of fish schools in a certain area during a certain season. The Poisson distribution has only one parameter: μ = mean = variance. The shape of the distribution depends on μ; for large μ, the distribution tends to be symmetric and close to the normal distri-

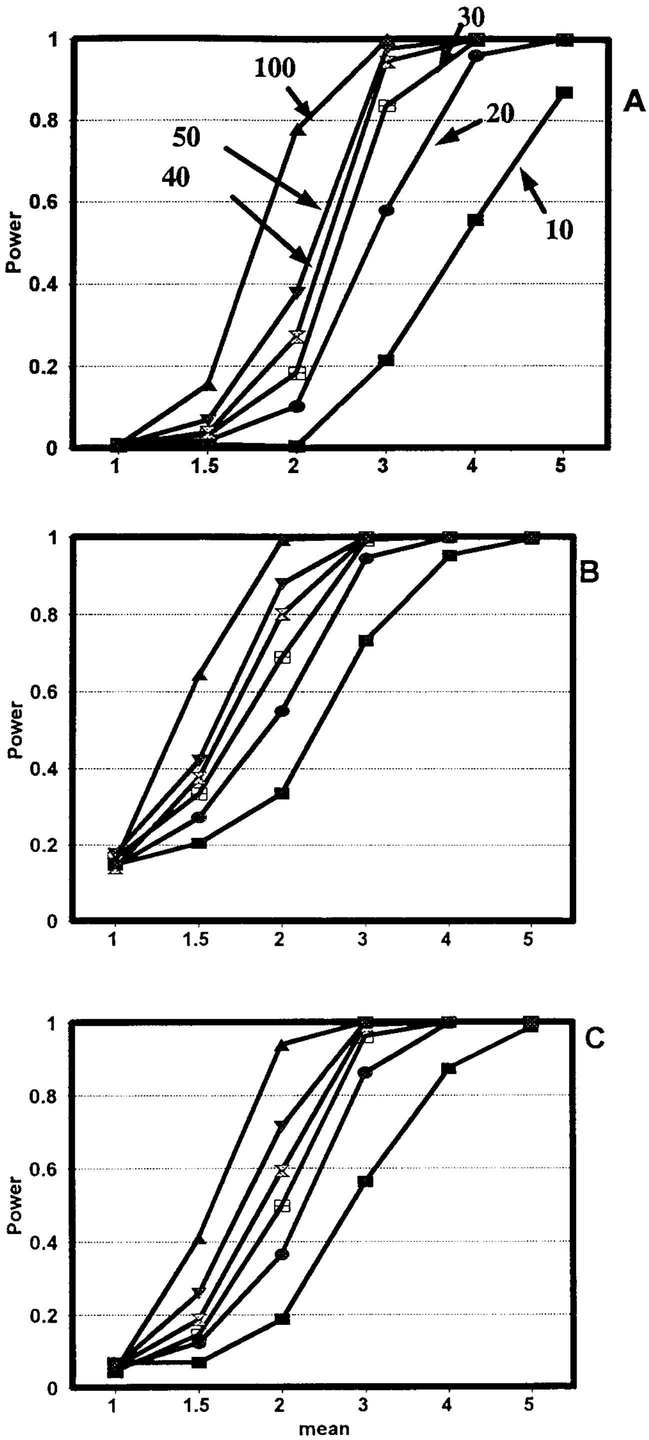

Figure 2. Power for procedures A, B, and C at different mean values (μ_2) compared to μ_1 = 1; σ = 2 for sample sizes ranging from 10 to 100 for normal distribution.

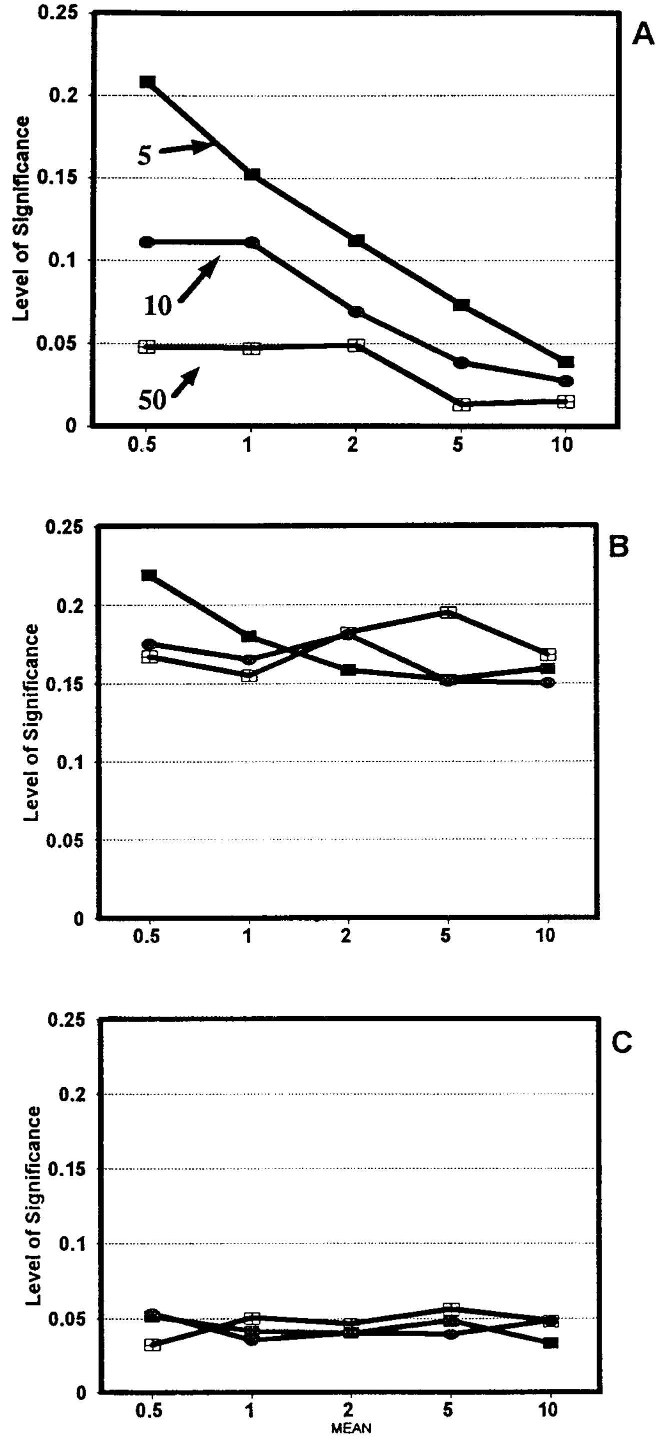

Figure 3. Level of significance (α) for procedures A, B, and C at different mean values for sample sizes 5, 10, and 50 for Poisson distribution.

bution. In the simulation, the α values were computed for μ = 0.5, 1, 2, 5, 10 and sample sizes 5, 10, and 50 (figure 3). Procedures A and B are sensitive to the sample size and mean values: the level of significance (α) decreases as the sample size and the mean values increase. But for procedure C, the α values (slightly less than 0.05) are independent of sample size and the mean values.

The α value averaged over all sample sizes and mean values was 0.06 for procedure A, 0.17 for procedure B, and close to 0.05 for procedure C (table 2 and figure 3).

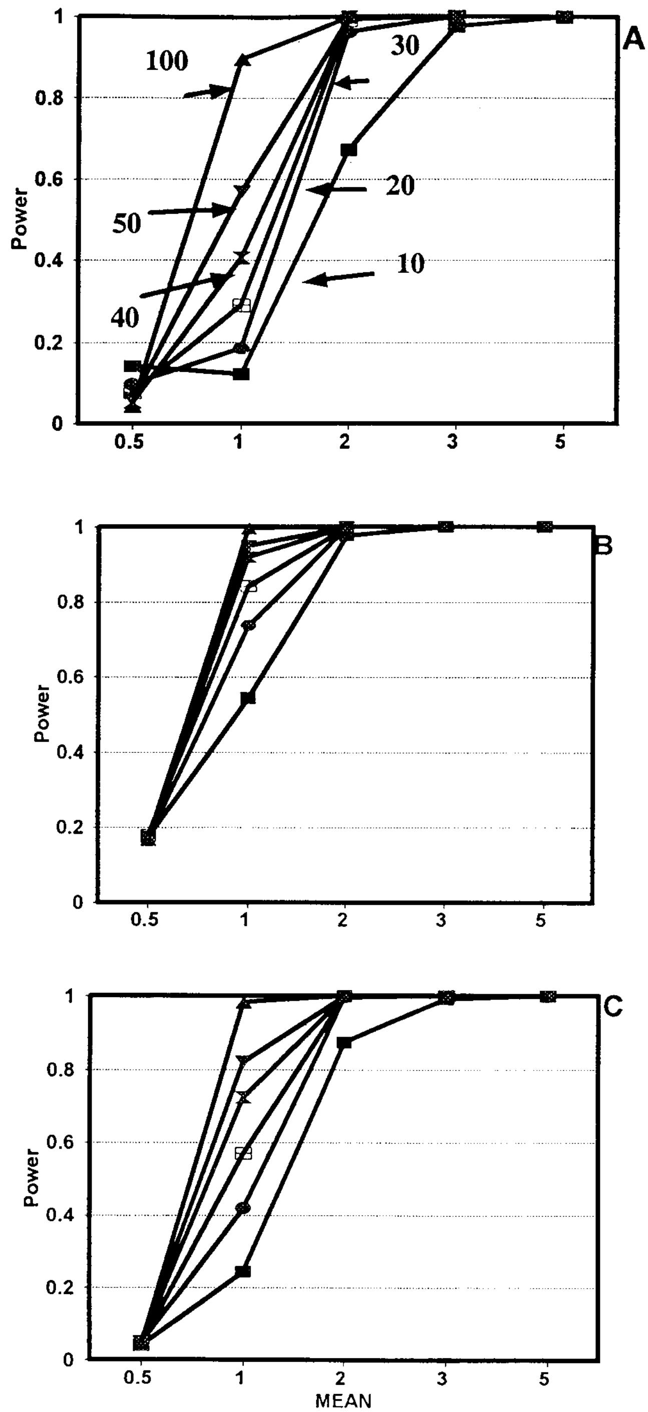

Figure 4. Power values for procedures A, B, and C at different mean values (μ_2) compared to μ_1 = 0.5 for sample sizes ranging from 10 to 100 for Poisson distribution.

Although the α values for sample sizes >10 are similar for procedures A and C, A is less powerful than C, regardless of sample sizes (figure 4). For example, for μ_2 = 1 compared to μ_1 = 0.5 for sample size 30, the power is 0.3 for procedure A, 0.84 for procedure B, and 0.57 for procedure C.

Gamma Distribution

The gamma distribution (equation A3) has two parameters (b,c) with mean = bc, and variance = b^2c. Without loss of generality, the scale parameter, b, was set at 1 (this is the exponential distribution). The shape of the gamma distribution tends to be symmetric as parameter c increases. Data were generated from six populations in which b = 1, and c = 0.5, 1, 1.5, 2, 2.5, or 3 compared to samples from one population with b = 1 and c = 3 (or μ_1 = 3).

For procedures A and C, the computed α values are independent of the mean values, with an average of 0.006 for A and 0.05 for C. For procedure B, the computed α values decrease as the sample size and the mean values increase (figure 5). The average of all the α values is 0.16 (table 2). Procedure B is the most powerful procedure, and A is the least powerful, regardless of sample size and mean values (figure 6). For example, for μ_2 = 2 compared to μ_1 = 3, with a sample size = 30, the power is 0.41 for procedure A, 0.85 for procedure B, and 0.72 for procedure C.

Lognormal Distribution

The lognormal distribution is often used to describe abundance per unit area (Meyers and Pepin 1990; Lo et al. 1992; equation A4). A random variable, y, follows lognormal if $x = \ln(y)$ is normal (μ, σ^2).

For each of three procedures, α was computed for $\mu_1 = \mu_2$ = 0, 0.5, 1, and 2; $\sigma_1 = \sigma_2$ = 1; and sample size n = 5, 10, 30, and 50. The α values are independent of sample size, with an average of 0.004 for procedure A, 0.19 for B, and 0.05 for C. For procedures A and B, the variation of α values increases as mean value increases (table 2, figure 7).

Procedure A is the least powerful of the three. For sample sizes <30, procedure B is most powerful. For sample sizes ≥30, procedure C is most powerful (figure 8). For example, for sample size = 30 and μ_2 = 1 compared to μ_1 = 0, the power is 0.63 for procedure A, 0.81 for procedure B, and 0.96 for procedure C.

CONCLUSIONS

Table 2 summarizes the simulation results for the level of significance and overall power for each of the four distributions.

The α value of procedure A is 0.005 for normal, gamma, and lognormal distributions. Thus procedure A probably leads to the conclusion that two means are the same even though they are different. But if the two CIs do not overlap, one can be much more than 95% sure that the population means are different. The α value of procedure A is not sensitive to sample size and mean values except for the Poisson distribution. For the Poisson,

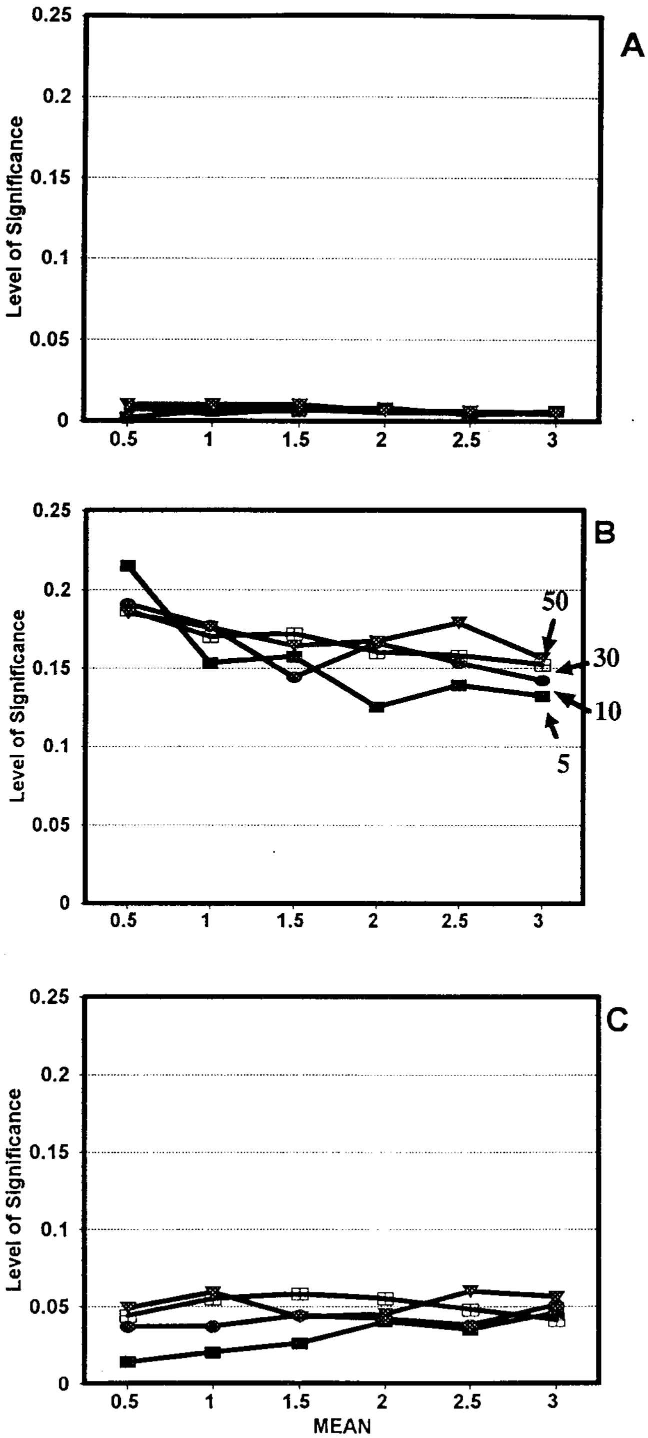

Figure 5. Level of significance (α) for gamma distribution (b = 1) for procedures A, B, and C at different mean values for sample sizes ranging from 5 to 50.

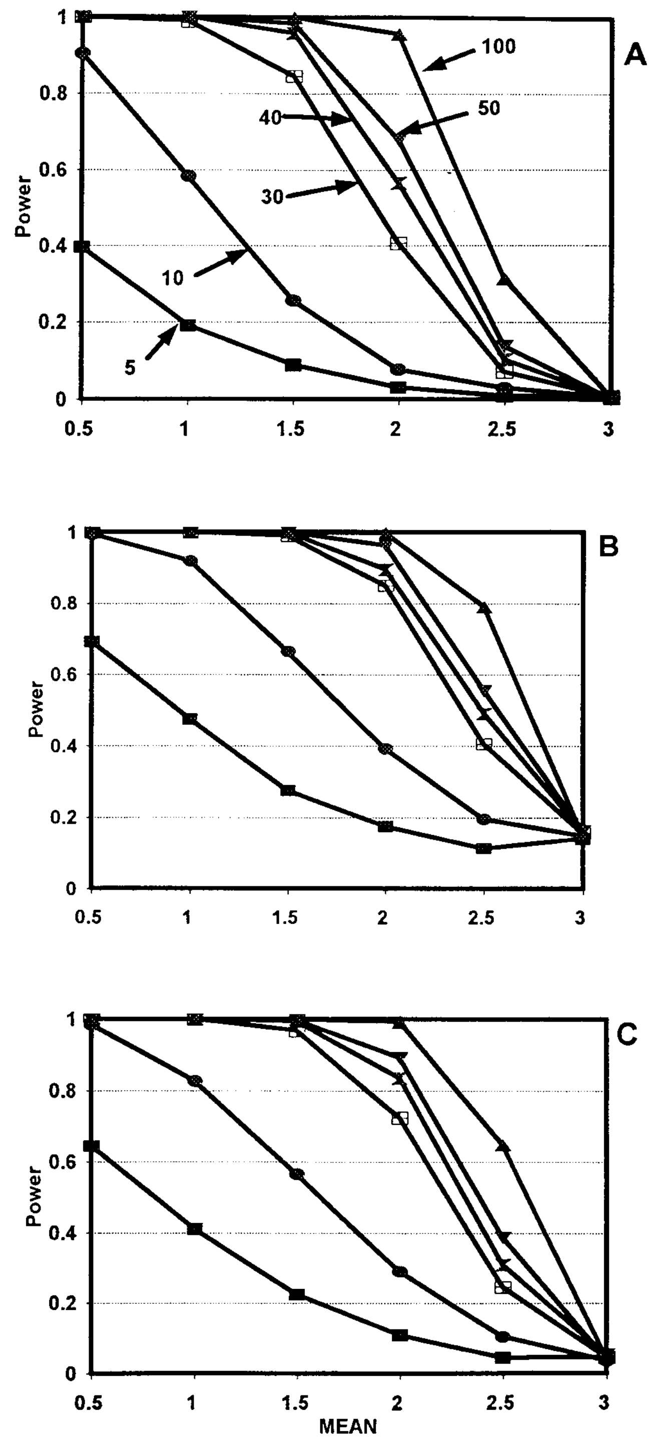

Figure 6. Power for procedures A, B, and C at different mean values (μ_2) compared to μ_1 = 3 for sample sizes ranging from 5 to 100 for gamma distribution.

the α is large for small sample size and small mean values; it decreases as the sample size or the mean values increase, with an overall average equal to 0.06 (figure 3).

The α value of procedure B is 0.17 averaged over all the distributions. Thus procedure B invites one to conclude that two population means are different when, in fact, they are the same. For the normal distribution, the α values differ for sample size < or > 10, and are not affected by the mean values. For Poisson distribution, α values decrease as the mean values increase for

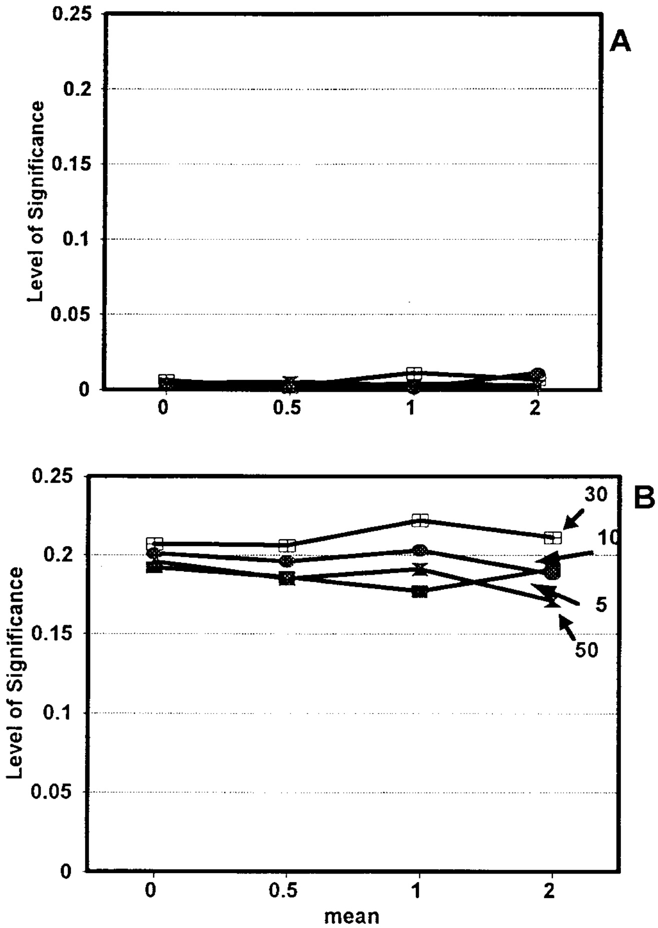

Figure 7. Level of significance (α) for procedures A and B at different mean values for sample sizes ranging from 5 to 50 for lognormal distribution (α values for procedure C are close to 0.05 and were not plotted.)

small sample size (n = 5). For gamma distribution, the α value of procedure B decreases as the sample size or the mean value increases, particularly for a sample size equal to 5.

Of the three procedures, B is most powerful, and A is least powerful if the population means are indeed different. The difference in the power values between procedures A and B results from the difference in α values; thus one should not conclude that procedure B is "better" than A based on power unless both A and B have the same α value. As expected, power increases with sample size and the difference between the means for all procedures.

Procedure C maintains the assumed confidence level for the four distributions when the sample size is greater than 10. Therefore, CI(d) is a robust procedure that can be used for these three nonnormal distributions if the sample size is moderate.

In the example of comparing the mean fecundity of anchovy, the fact that procedures A and B indicated different conclusions is not surprising now that we know

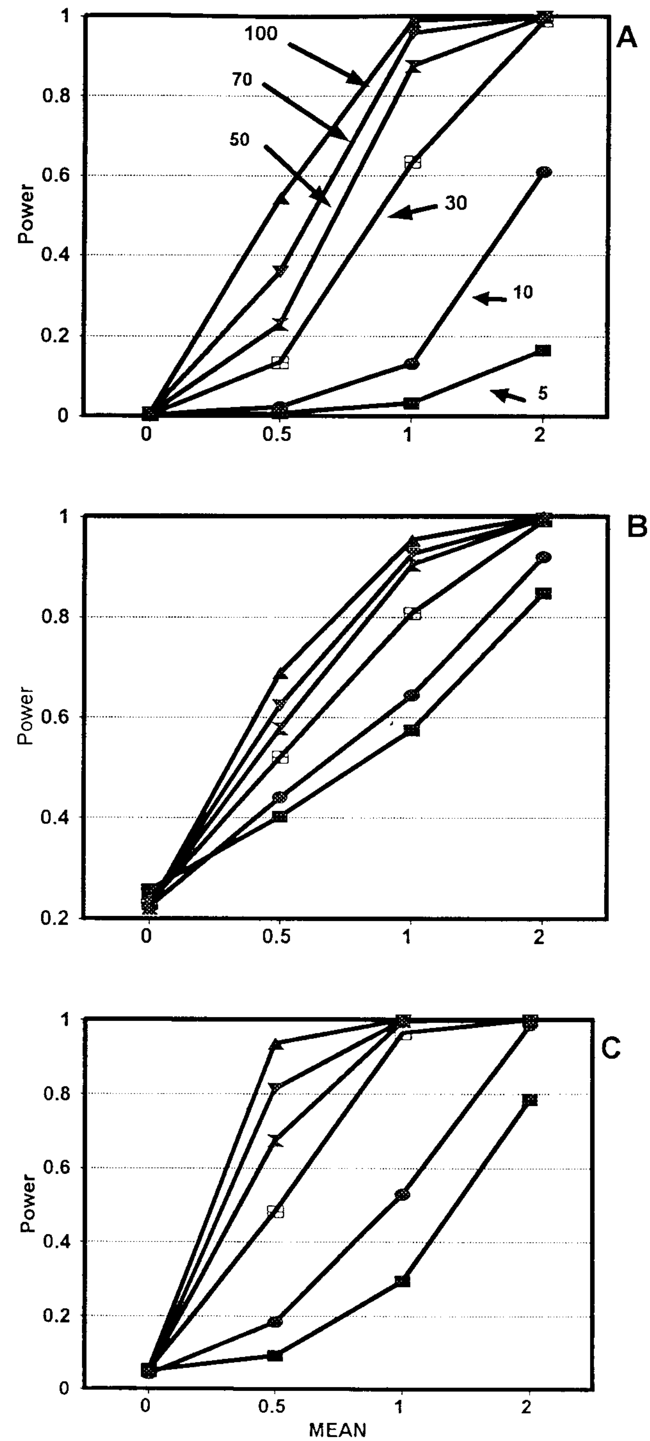

Figure 8. Power for procedures A, B, and C at different mean values (μ_2) compared to μ_1 = 0; σ = 1 for sample sizes ranging from 5 to 100 for lognormal distribution.

that procedure A has a low level of significance which leads to claiming that the two means are not statistically different, whereas procedure B does the opposite.

For the lognormal distribution, if the variances are not equal, variance should be included in the computation of the CI because the mean of y is equal to $\exp(\mu+\sigma^2/2)$. The CI should be one for $\exp(\mu+\sigma^2/2)$. The latter expression is in terms of the mean and variance of x.

For future research, it would be useful to

- obtain the level of significance of the original CI to achieve a 5% level of significance for procedures A and B for other distributions as reported by Nelson (1989) for the normal case
- investigate the effect of unequal sample sizes on the level of significance and power of the three procedures
- develop procedures for comparing two population means when the standard error is given but not the sample size.

ACKNOWLEDGMENTS

I thank Susan Picquelle of Northwest Fisheries Science Center, Seattle; Alfredo Julian, Larry Jacobson, and Paul Smith of Southwest Fisheries Science Center; and C. J. Park of San Diego State University for reading the manuscript and providing constructive comments.

LITERATURE CITED

Barr, D. R. 1969. Using confidence intervals to test hypotheses. J. Quality Technology 1:256–258.
Beal, S. L. 1989. Sample size determination for confidence intervals on the population mean and on the difference between two population means. Biometrics 45:969–977.
Casella, G., and C. Robert. 1989. Refining Poisson confidence intervals. Can. J. Statistics 17(1):45–57.
Douglas, J. B. 1993. Confidence regions for parameter pairs. Am. Statistician: 47(1):43–45.
Hunter, J. R., and R. Leong, 1981. The spawning energetics of female northern anchovy, *Engraulis mordax*. Fish. Bull., U. S. 79:215–230.
Land, C. E. 1988. Hypothesis tests and interval estimates. *In* Lognormal distributions, Crow and Shimizu, eds. Marcel Dekker, Inc.
Lo, N. C. H., L. D. Jacobson, and J. L. Squire. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Can. J. Fish. Aquat. Sci. 49:2515–2526.
Meyers, R. A., and P. Pepin. 1990. The robustness of lognormal-based estimators of abundance. Biometrics 46:1185–1192.
Nelson, L. S. 1989. Evaluating overlapping confidence intervals. J. Quality Technology. 21:140–141.
Peterman, R. M. 1990. Statistical power analysis can improve fisheries research and management. Can. J. Fish. Aquat. Sci. 47:2–14.
Solow, A. R., and J. H. Steele. 1990. On sample size, statistical power and the detection of density dependence. J. Anim. Ecol. 59:1073–1076.
Weerahandi, S. 1993. Generalized confidence intervals. J. Am. Stat. Assoc. 88(423):899–905.
Withers, C. S. 1991. Tests and confidence intervals for the shape parameter of a gamma distribution. Commu. Statist.-Theory Math. 20(8):2663–2685.
Zar, J. H. 1984. Biostatistical analysis. 2d ed. Englewood Cliffs, N. J.: Prentice-Hall, 718 pp.

APPENDIX

Density functions used in the simulation analysis

Normal

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)} \qquad \text{(A1)}$$

for $-\infty < x < \infty$; $0 < \sigma$, $-\infty < \mu < \infty$

Poisson

$$f(x;\lambda) = \lambda^x e^{-\lambda}/x!$$

for $x = 0, 1, 2, 3, \ldots ; 0 < \lambda$ (A2)

Gamma

$$f(x;b,c) = \frac{1}{\Gamma(c)\, b}\,(x/b)^{c-1} e^{-(x/b)} \qquad \text{(A3)}$$

for $0 < x$; $0 < b,c$

Lognormal

$$f(y;\mu,\sigma) = \frac{1}{y\,\sigma\sqrt{2\pi}}\, e^{-(\ln(y)-\mu)^2/(2\sigma^2)} \qquad \text{(A4)}$$

for $0 < y$; $0 < \sigma$, $-\infty < \mu < \infty$.