SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 25, No. 5, pp. 1441–1444, 2006

Printed in the USA
0730-7268/06 $12.00 + .00

Hazard/Risk Assessment
Short Communication

COMPARING MEDIAN LETHAL CONCENTRATION VALUES USING CONFIDENCE INTERVAL OVERLAP OR RATIO TESTS

MATTHEW W WHEELER,*† ROBERT M. PARK,† and A. JOHN BAILER,†‡
†National Institute for Occupational Safety and Health, 4676 Columbia Parkway MS-15, Cincinnati, Ohio 45226, USA
‡Center for Environmental Toxicology and Statistics, Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056, USA

(*Received* 16 *May* 2005; *Accepted* 23 *September* 2005)

Abstract—Experimenters in toxicology often compare the concentration–response relationship between two distinct populations using the median lethal concentration (LC50). This comparison is sometimes done by calculating the 95% confidence interval for the LC50 for each population, concluding that no significant difference exists if the two confidence intervals overlap. A more appropriate test compares the ratio of the LC50s to 1 or the log(LC50 ratio) to 0. In this ratio test, we conclude that no difference exists in LC50s if the confidence interval for the ratio of the LC50s contains 1 or the confidence interval for the log(LC50 ratio) contains 0. A Monte Carlo simulation study was conducted to compare the confidence interval overlap test to the ratio test. The confidence interval overlap test performs substantially below the nominal $\alpha = 0.05$ level, closer to $p = 0.005$; therefore, it has considerably less power for detecting true differences compared to the ratio test. The ratio-based method exhibited better type I error rates and superior power properties in comparison to the confidence interval overlap test. Thus, a ratio-based statistical procedure is preferred to using simple overlap of two independently derived confidence intervals.

Keywords—Median lethal concentration Significance testing Type I error rates Power Fieller's method

INTRODUCTION

Environmental toxicology experiments often examine the response of test organisms to concentrations of a toxicant. Survival is a common dichotomous response measured in these concentration–response studies. A summary of the potency of a toxicant, such as the median lethal concentration (LC50), can be calculated and then compared between different populations. For example, the impact of fluoranthene on fathead minnow survival may differ depending on whether the exposure occurs in the presence or absence of algae because of phototoxic effects (i.e., the presence of algae may mitigate the toxicity of fluoranthene in higher ultraviolet conditions). A formal comparison of fluoranthene in LC50s in response to different algal levels might be of interest. Such a comparison often is done by statistically testing the differences of the LC50 values. Statistical hypothesis tests commonly are employed to do this task.

In this situation, experimenters might compare confidence intervals (CIs) from two populations, and if these CIs overlap, the experimenters conclude that the populations are not statistically different. The use of the CI overlap test is driven by the simplicity and availability of CIs as part of standard statistical software, and this test is used in situations when the software does not perform appropriate tests on the statistics of interest. This is a flawed decision rule that results in overly conservative testing strategy. Schenker and Gentleman [1] provided a comprehensive analysis of the statistical inadequacy of using a CI overlap testing procedure. In particular, they concluded that ''[a CI overlap testing procedure] should not be used for formal significance testing unless the data analyst is aware of its deficiencies and unless the information needed to carry out a more appropriate procedure is unavailable'' [1, p 186]. Although the CI overlap method is flawed, thus making it more difficult to detect significant differences, its use frequently is reported within the literature. A nonexhaustive search of *Environmental Toxicology and Chemistry* found four articles, published between 2001 and 2004, that employed the CI overlap test for significance testing between two or more LC50s. Use of this test also has been reported in other journals and by the U.S. Environmental Protection Agency (U.S. EPA) for testing LC50 significance. The U.S. EPA toxicity reduction evaluation guidance document for municipal wastewater treatment plants [2] uses this decision rule to test the LC50 from two distinct populations, stating that ''[if the] LC50 had been 50% with confidence limits of 42 to 62%, the industrial wastewater would have been indicated as a possible source of toxicity based on the results of series 2 because the 95% confidence limits do not overlap . . . However, if dilution series 1 had been used, the industrial wastewater may not have been judged to be a toxic source because the confidence limits overlap'' [2, p 46]. Although this method frequently is used in the ecotoxicological literature, we are not aware of any formal comparison of this method against better methods in this environmental toxicology literature. Whereas the CI overlap test has been used for examining various parameters, including means and coefficients of variation, we investigate the behavior of the CI overlap test when comparing LC50 values based on the prevalence of its use in the literature.

In the present study, the CI overlap test is compared with a ratio test, which compares the ratio of the LC50s to 1 or the log(LC50 ratio) to 0. We do this by comparing the LC50s sampled from two distinct populations. More sophisticated techniques exist and are recommended for comparing concentration–response curves [3]. In the present study, however, we focus on comparing LC50s using the CI overlap test as well

* To whom correspondence may be addressed (mwheeler@cdc.gov).

as a ratio test, because more sophisticated tests, which compare concentration–response curves, do not necessarily test the hypothesis of the equality of the LC50s. A computer simulation study is used to conduct this comparison, and a motivating, real-data illustration of these two statistical testing methods is presented.

MATERIALS AND METHODS

The behavior of both the CI overlap test as well as the ratio test was studied through computer simulation. In this simulation, the true concentration–response for both populations was specified using a probit linear-regression model. This model describes the probability of the organism's death by a linear relationship that relates concentration to the cumulative distribution of the standard normal distribution. Mathematically, this is represented through the relationship $p = \Phi(\beta_0 + \beta_1 \cdot \text{concentration})$, where $\Phi(x)$ is the cumulative distribution function for a standard normal variate evaluated at x; that is, $\Phi(x) = \Pr(Z \leq x)$, where Z is distributed standard normally and $\Pr(\bullet)$ describes the probability an event occurs. For example, $\Phi(0) = 0.5$, and $\Phi(-1.96) = 0.025$. Because the LC50 is the concentration at which $\Phi(\beta_0 + \beta_1 \cdot \text{concentration}) = 0.5$, the LC50 can be expressed in terms of the probit regression parameters as $\text{LC50} = -\beta_0/\beta_1$. For experimental data containing two independent populations, the LC50 is estimated using output from generalized linear model (GLiM) software, such as the procedure PROC PROBIT in SAS (Ver 9.1; SAS Institute Cary, NC, USA) or the function glm in R (version 2.0.1, R Foundation, Vienna, Austria), and the equality of the groups is tested using the following constructions of the CI overlap and the LC50 ratio test.

In certain circumstances, tests like the likelihood ratio test can compare LC50 values and perform at the nominal level, but such a test is dependent on the assumption that either the backgrounds (β_0) or the slopes (β_1) are the same for both concentration–response curves. (For information on such tests, see Collett [4].) It is possible to have the same LC50 derived from concentration–response curves, which have different backgrounds and concentration–responses. Therefore, we focused on the CI overlap and the LC50 ratio test, because they directly test the equality of LC50s derived from separate populations.

CI overlap test

The output from most GLiM software includes estimates of β_0 and β_1 along with the approximated variance–covariance matrix. Given these estimates, the LC50 can be estimated, and approximate large-sample $(1 - \alpha)$% CIs can be constructed using Fieller's Theorem [5]. The formula

$$\widehat{\text{LC50}} + \frac{\gamma}{1-\gamma}\left(\widehat{\text{LC50}} + \frac{c_{12}}{se^2[\hat{\beta}_1]}\right) \pm \frac{z_{\alpha/2}}{(1-\gamma)|\hat{\beta}_1|}\left\{se^2[\hat{\beta}_0] + 2c_{12}\widehat{\text{LC50}} + se^2[\hat{\beta}_1]\widehat{\text{LC50}}^2 - \gamma\left(se^2[\hat{\beta}_0] - \frac{c_{12}^2}{se^2[\hat{\beta}_1]}\right)\right\}^{1/2} \quad (1)$$

describes this interval, where $se^2[\hat{\beta}_0]$ and $se^2[\hat{\beta}_1]$ represents the square of the standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively; γ represents the quantity $z_{\alpha/2}^2\, se^2[\hat{\beta}_1]/\hat{\beta}_1^2$; and c_{12} is the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$. Using the CI described above, the statistical hypothesis test for the alternative hypothesis H_a : $\text{LC50}^a \neq \text{LC50}^b$ (the LC50 differs between populations a and b) is carried out by checking if the CIs overlap. If the CIs do not overlap, the null hypothesis H_0 (the LC50 is the same between populations a and b) is rejected in favor of the alternative; otherwise, one fails to reject the null hypothesis.

Table 1. Experimental data[a]

Fluoranthene concentration (μg/L)	Algal concentration	
	0.7×10^5 cells	1.5×10^5 cells
5	0/23, 0/19, 0/20	0/22, 0/25, 0/24
10	3/21, 4/20, 3/21	0/20, 0/27, 2/27
20	15/23, 14/19, 17/23	13/20, 12/18, 18/22
30	23/23, 20/20, 20/20	20/21, 17/19, 18/22

[a] Data provided by James T. Oris and Eun-ah Cho (Miami University, Oxford, OH, USA). Proportions represent fathead minnow larvae mortality after 4 h of exposure to fluoranthene.

LC50 ratio test

It also is possible to test the equality of the LC50s by examining the ratio of LC50s. We define this ratio as $\zeta = \text{LC50}^a/\text{LC50}^b$ for the LC50s from the independent populations a and b, respectively. Given $\hat{\zeta}$, the estimate of ζ, and $\text{var}(\hat{\zeta})$, the variance of the estimated parameter, which can be approximated using the delta method [6], an approximate $100(1 - \alpha)$% CI for ζ can be constructed. If the CI for the LC50 ratio contains 1, we fail to reject the hypothesis that the population LC50s are the same. Because the sampling distribution of ratio statistics, such as the $\widehat{\text{LC50}}$ and the above-defined $\hat{\zeta}$, have a tendency to be skewed, a log transformation often is used to normalize the distribution of these statistics. Consider $\log(\zeta) = \log(\text{LC50}^a) - \log(\text{LC50}^b)$, the log of the ratio of the LC50s. Assuming the independence of populations a and b, the standard error (SE) of $\log(\hat{\zeta})$ is estimated using the formula

$$\widehat{\text{SE}}[\log(\zeta)] = \sqrt{\frac{se^2[\hat{\beta}_0^a]}{(\hat{\beta}_0^a)^2} + \frac{se^2[\hat{\beta}_1^a]}{(\hat{\beta}_1^a)^2} + \frac{se^2[\hat{\beta}_0^b]}{(\hat{\beta}_0^b)^2} + \frac{se^2[\hat{\beta}_1^b]}{(\hat{\beta}_1^b)^2} - \frac{2c_{12}^a}{\hat{\beta}_0^a\hat{\beta}_1^a} - \frac{2c_{12}^b}{\hat{\beta}_0^b\hat{\beta}_1^b}} \quad (2)$$

where the parameters are defined as in Equation 1. If the CI for the $\log(\zeta)$ contains 0, we fail to reject the hypothesis that the LC50s are equal. A Z test also can be constructed from this procedure, where $z = \log(\hat{\zeta})/se[\log(\hat{\zeta})]$. We call this test the LC50 ratio test.

It is important not to confuse the LC50 ratio test with the ''ratio test'' studied for similar purposes by Payton et al [7]. In their test, the term *ratio* refers to the ratio formed from regression coefficients when computing the LC50, not to the ratio of LC50s derived from two independent populations. (For more information on the testing procedure that they used, see Robertson and Preisler [8].)

Fluoranthene data

We illustrate the difference between the CI overlap test and the ratio method in comparing LC50s in an experiment described in Table 1, in which fathead minnow larvae are exposed to fluoranthene, a polycyclic aromatic hydrocarbon (PAH; J.T. Oris and Eun-ah Cho, Miami University, Oxford, OH, USA, unpublished data). Because studies have shown increased toxicity of PAHs in the presence of ultraviolet radiation, we compared the acute toxicity of fluoranthene between two groups

Table 2. Estimated concentration–response curves of fathead minnow larvae mortality after 4 h of exposure to fluoranthene[a]

Algal concentration (cells)	Intercept	Standard error	Slope	Standard error	Covariance	LC50	95% Confidence interval
0.7×10^5	−5.26	0.633	0.317	0.0362	−0.0215	16.6	(15.2, 18.0)
1.5×10^5	−5.58	0.653	0.293	0.0339	−0.0210	19.0	(17.6, 20.5)

LC50 ratio test	LC50 ratio test statistic	Standard error	Test statistic	*p*
	0.138	0.0559	2.46	0.0136

[a] The median lethal concentration (LC50) as well as corresponding 95% confidence limits are reported and compared to the LC50 ratio test.

of fathead minnow larvae living in environments where different algal densities result in different ambient ultraviolet levels. For illustration purposes, we look at the experiments having algal concentrations of 0.7×10^5 and 1.5×10^5 cells. For a given algal density, four groups of approximately 20 larvae were exposed to fluoranthene concentrations of 5, 10, 20, and 30 μg/L for 4 h, and the proportion of fish dying was the measured outcome for each group. This experiment was repeated three times, giving multiple replications per data point. A concentration–response curve was then fit using probit regression. These fits are described in Table 2.

Estimated fluoranthene LC50s (95% CIs) were 16.6 (15.2, 18.0) and 19.0 (17.6, 20.5) μg/L for the algal concentrations of 0.7×10^5 and 1.5×10^5 algal cells, respectively. The CIs for these values overlap, and the CI overlap test would suggest no difference in fluoranthene-associated toxicity in fathead minnow larvae exposed to fluoranthene at the above algal densities. In contrast, the LC50 ratio test provided a *p* value of 0.014, which is associated with an observed *Z* of 2.46, allowing one to conclude the LC50s differ under the two algal conditions. This raises the question of what is behind the discrepancy in these two statistical methods.

Simulation design

Because the LC50 is defined by the parameters β_0 and β_1, we proceeded in simulating data by first specifying these parameters and then generating random binomial experiments based on these parameters. Specifically, three concentration–response patterns were specified that reflect different β_0 and β_1 choices (Fig. 1). These patterns, which represent all possible experimental outcomes in which the concentration–response curves differ, included a concentration–response curve that differed only in background response (β_0), a curve that differed only in the slope of the concentration–response (β_1), and a curve that differed in both intercept and slope. For all three patterns, the concentration–response parameters of the first population were fixed at $\beta_0 = -1.75$ and $\beta_1 = 0.5$; this translated to a background response of $p = \Phi(-1.75) = 0.04$ for unexposed (concentration = 0) organisms. For the different background response pattern, the slope of the second population was fixed at $\beta_1 = 0.5$, and the β_0 term was varied between −1.75 and −0.75 in increments of 0.033. For the different slope pattern, the second population had the β_0 term fixed at −1.75, and the slope (β_1) was varied between 0.5 and 0.8 in increments of 0.01. Finally, for the pattern in which both the background and concentration–response differed, the second population had the β_0 term fixed at −1 and the β_1 term varied between 0.29 and 0.59 in increments of 0.01.

Given the concentration–response patterns, the simulation proceeded by sampling at the concentrations of {0, 1, 2, 3, 4, 4.5}, with 30 observations per concentration. Setting $\alpha = 0.05$, and given each unique value of β_0 and β_1, 2,000 simulated experiments were generated, and the CI overlap test and the LC50 ratio test were calculated. The operating characteristics of these tests (false positive/type I error rates and test sensitivity/power) based on the overlap test as well as the LC50 ratio test were examined. We determined the number of simulations so that the margin of error when estimating rejection rates (e.g., type I error rates or power) was approximately 0.01 [9], which leads to accurate estimates of the true type I error rate. All simulations were conducted using the R statistical programming environment [10].

RESULTS

Table 3 reports the observed type I error rate for the two tests across the three concentration–response conditions. The CI overlap provided an observed type I error rate of between 0.004 and 0.005 for the three conditions. The type I error rate for the ratio test was closer to that of the nominal value of 0.05 and ranged between 0.035 and 0.044 across the simulation conditions.

The conservative nature of the CI overlap test would reduce the test's ability to detect differences when, in fact, they did occur. Figure 2 graphically displays the probability that a given test detects a difference between LC50s for the different intercept condition as a function of the true difference in LC50s.

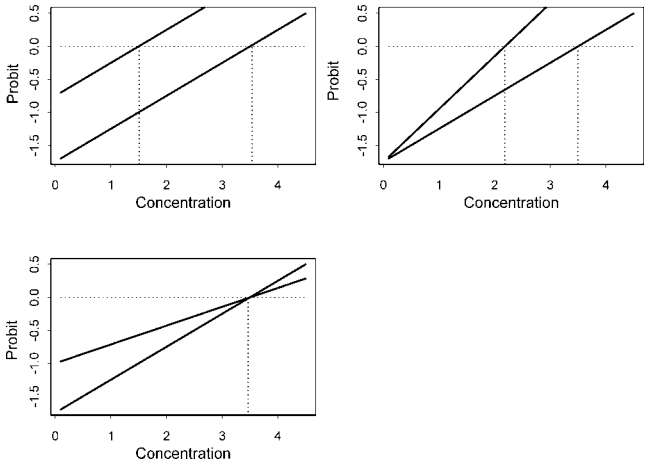

Fig. 1. Graphical representation of the three concentration–response functions used in the simulations. The horizontal line corresponds to the probit score associated with a 50% probability of response. Vertical lines represent the true median lethal concentration (LC50) for the concentration–response curve.

Table 3. Observed type I error rate for the three simulations[a]

Simulation condition	Confidence interval overlap	LC50 ratio test
Different background rate	0.005	0.044
Different slope	0.004	0.041
Different slope and background	0.005	0.035

[a] The nominal type I error rate for the confidence interval overlap test and median lethal concentration (LC50) ratio test was set at $\alpha = 0.05$.

For simulated concentration–response curves, the power of the ratio test was greater than that of the CI overlap test by 20 to 30% for most LC50 differences between populations. The results show that to accurately classify a true difference 50% of the time, the CI overlap test requires that the actual difference between the LC50s be approximately 40% greater than what would be required using the ratio test. The LC50 ratio test obtains a power of approximately 80% for a true LC50 difference of $LC50^a - LC50^b = 0.8$ units, whereas the CI overlap test only had a 50% chance of detecting this difference. A similar pattern was observed for the other simulation conditions (not shown).

DISCUSSION

The CI overlap test provides a simple method of testing LC50s, but the test also provides less sensitive inference when comparing these values. In contrast, the LC50 ratio test provides an equally straightforward method of testing two LC50s while providing more sensitive inference. It is worth noting that the LC50 ratio test, like the CI overlap test, does not require access to the data; rather, it relies on the reported $\hat{\beta}_0$ and $\hat{\beta}_1$ as well as their estimated variance and covariance, which are values that could easily be included in any manuscript. If parameter estimates are available to the experimenter, we argue that the LC50 ratio test be used over the CI overlap test and that, in general, the CI overlap test should be used only when no alternative test exists. This conclusion is based on the observed error rate of approximately 0.005 when the nominal type I error rate was set at $\alpha = 0.05$, compared to an observed error rate of approximately 0.04 for the LC50 ratio test. The conservative nature of the CI overlap test also decreases the probability that a significant difference is found when, in fact, a difference exists between the populations. Thus, comparison of multiple parameters should be incorporated explicitly in the construction of hypothesis tests and CIs. Further summary figures and tables, which report the population statistics as well as their corresponding 95% CIs, should annotate significant effects based on formal hypothesis testing; this may prevent one from drawing inappropriate conclusions when CIs overlap. Finally, as noted earlier, more sophisticated statistical testing options are available for comparing concentration–response curves; however, if one desires to compare toxicity using the derived LC50s, we recommend use of the LC50 ratio test over the CI overlap test.

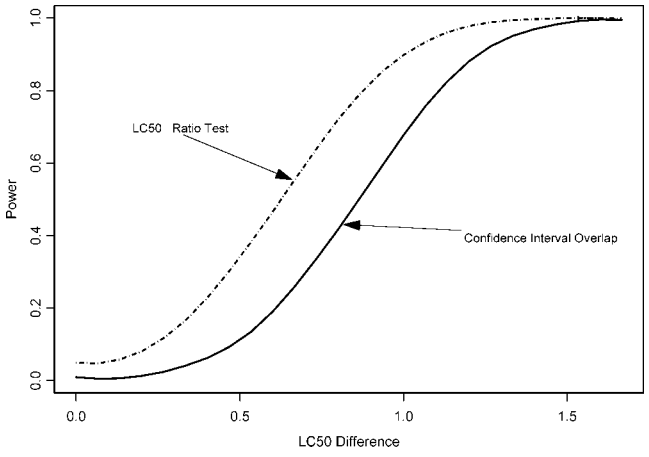

Fig. 2. Estimated power of the confidence interval overlap test and medial lethal concentration (LC50) ratio test in finding a significant difference between LC50 values (based on 2,000 simulations at each specified LC50 difference). The figure depicts a smoothed loess line fitted to the simulation under the condition that only β_0 differs in $p = \Phi(\beta_0 + \beta_1 \cdot \text{concentration})$, where β_0 represents the intercept and β_0 represents the slope term in the equation between two experimental groups. The LC50 difference is described as the true arithmetic difference between the two groups.

Acknowledgement—We thank James Deddens, James T. Oris, and Marc Greenberg along with the editor and two anonymous referees for their comments on earlier versions of this manuscript. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

REFERENCES

1. Schenker N, Gentleman JF. 2001. On judging the significance of differences by examining by examining the overlap between confidence intervals. *The American Statistician* 55:182–186.
2. U.S. Environmental Protection Agency. 1999. Toxicity reduction evaluation guidance for municipal wastewater treatment plants. EPA 833B-99/002. Final Report. Office of Wastewater Management, Washington, DC.
3. Oris JT, Bailer AJ. 1997. Equivalence of concentration–response distributions in aquatic toxicology: Testing and implications for potency estimation. *Environ Toxicol Chem* 16:2204–2209.
4. Collett D. 2003. *Modeling Binary Data*, 2nd ed. CRC, Boca Raton, FL, USA.
5. Buonaccorsi J. 2002. Fieller's theorem. In El-Saarawi A, Piegorsch W, eds, *Encyclopedia of Environmetrics*, Vol 2. John Wiley, New York, NY, USA, pp 773–774.
6. Cox C. 2002. Delta method. In El-Saarawi A, Piegorsch W, eds, *Encyclopedia of Environmetrics*, Vol 1. John Wiley, New York, NY, USA, pp 488–489.
7. Payton ME, Greenstone MH, Schenker N. 2003. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *J Insect Sci* 3:1–6.
8. Robertson JL, Preisler HK. 1992. *Pesticide Bioassays with Arthropods.* CRC, Boca Raton, FL, USA.
9. Thompson SK. 1992. *Sampling*. John Wiley, New York, NY, USA.
10. Venables WN, Smith DM. 2002. *An Introduction to R.* Network Theory, Bristol, UK.