Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance?

Authors: Payton, Mark E., Greenstone, Matthew H., and Schenker, Nathaniel

Source: Journal of Insect Science, 3(34) : 1-6

Published By: Entomological Society of America

URL: https://doi.org/10.1673/031.003.3401

Payton ME, Greenstone MH, Schenker N. 2003. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? 6pp. *Journal of Insect Science*, 3:34, Available online: insectscience.org/3.34

Journal of Insect Science

insectscience.org

Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance?

Mark E. Payton[1], Matthew H. Greenstone[2], and Nathaniel Schenker[3]

[1] *Department of Statistics, 301 MSCS Building, Oklahoma State University, Stillwater, OK 74078-1056*
[2] *U.S. Department of Agriculture, Agricultural Research Service, Insect Biocontrol Lab, Bldg. 011A, Rm. 214, BARC-West, 10300 Baltimore Avenue, Beltsville, MD 20705*
[3] *Office of Research and Methodology, National Center for Health Statistics, 3311 Toledo Road, Room 3209, Hyattsville, MD 20782*
mpayton@okstate.edu

Received 20 June 2003, Accepted 2 October 2003, Published 30 October 2003

Abstract

We investigate the procedure of checking for overlap between confidence intervals or standard error intervals to draw conclusions regarding hypotheses about differences between population parameters. Mathematical expressions and algebraic manipulations are given, and computer simulations are performed to assess the usefulness of confidence and standard error intervals in this manner. We make recommendations for their use in situations in which standard tests of hypotheses do not exist. An example is given that tests this methodology for comparing effective dose levels in independent probit regressions, an application that is also pertinent to derivations of LC_{50}s for insect pathogens and of detectability half-lives for prey proteins or DNA sequences in predator gut analysis.

Keywords: Confidence intervals, detectability half-life, effective dose, LD_{50}, lethal dose, predator gut analysis, probit analysis, standard error intervals

Introduction

Scientists often express the results of experiments and observations by the use of means along with a measure of variability. For example, an insect ecologist or physiologist might have an experiment involving a number of treatments, and a systematist might have a sample of morphological measurements of the same character from a series of species. The results of each treatment or set of measurements are represented by a mean ± standard deviation or estimated standard error. Some refer to the interval using the estimated standard error, or standard deviation divided by the sample size, as a standard error interval. This approach is very useful in that it provides the reader with information regarding the measure of central tendency (mean) along with some idea of the variability (standard error) realized in the experiment. However, researchers sometimes fall into a trap by trying to use such results as a substitute for a hypothesis test. When the mean ± the estimated standard error for one treatment doesn't overlap with the corresponding interval for another treatment, the researcher might be tempted to conclude that the treatment means are different. This is a dangerous practice since the error rate associated with this comparison is quite large, with differences between equal treatment means declared significant more often than desired (Payton *et al*., 2000). Some will counter this problem by performing 95% confidence intervals and checking for overlap. However, this practice goes to the other extreme and creates extremely conservative comparisons, making it difficult to detect significant differences in means.

Occasionally a situation arises in which a test for the equality of two population parameters is needed but none exists, or at least not one that is easily applied. An example of this is testing the difference between coefficients of variation of random samples from two populations. This poses a unique testing problem since the technique for estimating the standard error associated with the coefficient of variation is not widely known, and thus a measure of variability is often not available for performing a test. Tests for coefficients of variation do exist (e.g., Gupta and Ma, 1996; Wilson and Payton, 2002), but they are somewhat complex and require specialized computer code that is not readily available. An approach one might take in this situation would be to calculate a confidence interval for the coefficient of variation from each sample, then declare them significantly different if the intervals do not overlap (relatively straight-forward methods for calculating confidence intervals for coefficients of variation are discussed in Vangel (1996) and Payton (1996)). The primary question becomes: What size of confidence interval should one set in this scenario to assure that the resulting test is at an acceptable error rate, say, 5%?

Previous work on the topic of hypothesis testing includes Payton *et al*. (2000) and Schenker and Gentleman (2001), both of

which explore the error rates observed when checking for overlap of standard error bars or confidence intervals in a testing situation. Browne (1979) explored such use of these intervals in what he called "visual tests" and how they related to tests of means. Goldstein and Healy (1995) proposed methodology that adjusted comparisons based on graphical representations of confidence intervals to attain a desired average type I error rate. We build on these articles and explore further the examination of overlap between confidence intervals or standard error intervals in comparing two population parameters. We discuss adjustments to be made in the event such a procedure needs to be used. We also extend this work to comparing lethal dose estimates and analogous response estimates from two independent probit regressions with the use of adjusted fiducial limits, which has applications to insect pathology and arthropod predation studies.

Confidence intervals and corresponding adjustments for testing hypotheses

Let's consider the situation of having random samples from two normally distributed populations. Let $\overline{Y}_1$ and $\overline{Y}_2$ be the sample means and let S_1 and S_2 be the sample standard deviations calculated from these random samples of size *n1* and *n2*. What we wish to do in this scenario is demonstrate the consequences of checking for overlap between unadjusted confidence intervals or standard error intervals to test hypotheses about the difference between two population means.

To calculate (1-α)100% confidence intervals for the mean, the formula is

$$(1) \qquad (\overline{Y}_i - t_{\alpha/2,ni-1}(S_i/\sqrt{ni}), \overline{Y}_i + t_{\alpha/2,ni-1}(S_i/\sqrt{ni}))$$

This formula is calculated for the samples from both populations (i.e., for i = 1 and 2). We can calculate the probability that the two intervals will overlap. This involves creating a probability expression for the situation in which the upper confidence limit from either sample is contained within the confidence limits of the other sample. If you allow the variable "*A*" to denote these intervals overlapping, this expression is given by

$$(2) \quad \Pr(A) = 1 - \Pr(not\_A) = 1 - \Pr[\overline{Y}_1 + t_{\alpha/2,n1-1}(S_1/\sqrt{n1}) < \overline{Y}_2 - t_{\alpha/2,n2-1}(S_2/\sqrt{n2})] - \Pr[\overline{Y}_2 + t_{\alpha/2,n2-1}(S_2/\sqrt{n2}) < \overline{Y}_1 - t_{\alpha/2,n1-1}(S_1/\sqrt{n1})]$$

If $n1 = n2 = n$, formula (2) simplifies to

$$(3) \qquad \Pr(A) = \Pr\left[\frac{n(\overline{Y}_1 - \overline{Y}_2)^2}{S_1^2 + S_2^2} < F_{\alpha,1,n-1}\frac{(S_1 + S_2)^2}{S_1^2 + S_2^2}\right]$$

The details of the algebraic manipulation leading to the above formula are given in Payton *et al.* (2000). One should note that the *F* value arises by squaring the *t* value in the original formula.

If the two populations being sampled are identical normal populations (i.e., same means and variances), the quantity $\frac{n(\overline{Y}_1 - \overline{Y}_2)^2}{S_1^2 + S_2^2}$ can be modeled with the *F* distribution with 1 and *n*-1 degrees of freedom. Therefore, the probability that the two intervals overlap can be denoted by

$$(4) \qquad \Pr(Intervals\_overlap) = \Pr\left[F_{1,n-1} < F_{\alpha,1,n-1}\left(1 + \frac{2S_1S_2}{S_1^2 + S_2^2}\right)\right]$$

A large-sample version of the above statement can be derived (again if one assumes that the two populations are the same):

$$(5) \qquad \Pr(Intervals\_overlap) \cong \Pr\left[|Z| < z_{\alpha/2}\sqrt{2}\right]$$

where $z_{\alpha/2}$ is the upper 100α/2 percentile of a standard normal variate (Z). The normal variate Z is used as the large-sample approximation for the square root of an *F*-distributed variate, and the parenthetical expression in (4) is replaced by the value *2* under the assumption of equality of population standard deviations. This result will illustrate the problem associated with checking for overlap between 95% confidence intervals as a testing device. If you set α = 0.05 and generate 95% confidence intervals, then the approximate probability of overlap can be calculated from expression (5) as

$$(6) \qquad \Pr(Intervals\_overlap) \cong \Pr[-2.77 < Z < 2.77] = 0.994$$

In other words, the 95% confidence intervals will overlap over 99% of the time. The consequences of using 95% confidence intervals should be evident. If you compare these intervals with the expectation of mimicking an α = 0.05 test, what you actually would be doing is performing a test with a much too conservative type I error rate. In other words, the 95% intervals are too wide, resulting in a procedure that declares differences at a proportion much less than the desired α = 0.05 rate.

We can make similar calculations regarding the use of standard error intervals, or intervals calculated by adding and subtracting the estimated standard error from the mean. Often researchers report their results in this fashion, and many times they will place standard error bars on graphs or figures. The easy trap to fall into, however, is thinking that because standard error bars associated with two means don't overlap, these means must be significantly different.

The large-sample probability of standard error intervals overlapping when the two populations are identical can be easily found by using expression (5) and replacing $z_{\alpha/2}$ with 1. Therefore

$$(7) \qquad \Pr(se\_intervals\_overlap) \cong \Pr\left[|Z| < \sqrt{2}\right]$$

This probability is equal to 0.843. Thus, examining overlap between standard error intervals to test hypotheses regarding equality of means would be akin to performing a test with a type I error rate of about 15% or 16%.

Schenker and Gentleman (2001) showed that for general estimation problems, the interval overlap method tends to be the most conservative when the (true) standard errors are equal. They found that for large samples, the Type I error rate when comparing the overlap of 100(1-γ)% confidence intervals is

$$(8) \qquad 2\Pr[Z < -z_{\gamma/2}(1+k)/\sqrt{1+k^2}]$$

where k is the ratio of standard errors. An analogous expression for the case of estimating means was given in Goldstein and Healy (1995). Replacing k with the value of 1 (i.e., assuming the standard errors are equal) will yield a multiplier for the z value in the probability statement of $\sqrt{2}$, which corresponds to the value given in expression (5).

Tables 1 and 2, based on expression (7), illustrate the relationship of standard error ratios to the likelihood of confidence intervals or standard error intervals overlapping. The data illustrate that the probabilities of overlap decrease as the standard errors become less homogeneous.

We can use equation (7) to guide us in adjusting the confidence limits for the intervals to achieve a more desirable error rate. For a given ratio of standard errors, k, setting equation (7) equal to a desired error rate of α=0.05 and solving for γ yields the correct large-sample confidence level that should be used for the individual intervals. For example, assuming equal standard errors ($k = 1$) yields $\gamma = 0.166$. In other words, if you wish to use confidence intervals to test equality of two parameters when the standard errors are approximately equal, you would want to use approximately 83% or 84% confidence intervals. A similar suggestion for the case of estimating means was made in Goldstein and Healy (1995). The sizes of the individual confidence intervals necessary to perform a 0.05 test grow as the standard errors become less homogeneous, as illustrated in Table 3.

A researcher will rarely know the true ratio of standard errors. One might estimate it with sample values. Of course, the method of comparing intervals is most useful for cases in which estimates for standard errors are not available. A possible approximation to the ratio of standard errors could be the ratio of the square roots of the two sample sizes, since the standard error of an estimate tends to be inversely proportional to the sample size.

We performed a simulation study to illustrate the calculations given above and to see how well the large-sample results apply to situations with small to moderate samples. Ten thousand pairs of independent random samples were generated from a standard normal distribution using PC SAS (SAS Inst., Cary, NC, 1996) Version 8.2. We varied the sample sizes from $n = 5$ to $n = 50$. Three intervals were constructed for each random sample: mean ± estimated standard error, 95% and 84% confidence intervals for the mean.

Results of the computer simulation are given in Table 4. The columns of the table record the proportion of times that the intervals for the pairs of random samples overlap. For instance, in the case where the sample size was 10, the proportion of the 10,000 iterations in which the two intervals constructed by the 95% confidence intervals overlapped was 0.995. The proportion of the 10,000 trials in which the two 84% confidence intervals overlapped for the $n = 10$ case was 0.949.

These simulation results validate much of the work done in the previous section. In particular, we have demonstrated that examining the overlap of 95% confidence intervals to test hypotheses is much too conservative. Likewise, using standard error intervals will produce the opposite effect. Another important outcome is the results of using 84% confidence interval methodology (when the true standard errors are equal). The adjusted intervals seem to work well for all sample sizes.

Table 1. Large-sample probability of overlap of 95% confidence intervals under the null hypothesis

Ratio of standard errors	1	2	3	4	5
Probability of overlap	0.994	0.991	0.987	0.983	0.979

Table 2. Large-sample probability of overlap of standard error intervals under the null hypothesis

Ratio of standard errors	1	2	3	4	5
Probability of overlap	0.843	0.82	0.794	0.775	0.761

Table 3. Large-sample confidence levels of individual intervals that yield a probability of overlap of 0.95

Ratio of standard errors	1	2	3	4	5
Confidence level (%) for individual intervals	83.4	85.6	87.9	89.4	90.4

Comparing effective dosages from independent probit regressions

Binary regression is useful in experiments in which the relationship of a response variable with two levels to a continuous explanatory variable is of interest. These are often referred to as dose-response models. Sometimes researchers are interested in estimating the dose that is needed to produce a given probability. For example, what insecticide dose is needed to provide an estimated probability of 0.95 for killing an insect? An estimate of this dose is important because using more than is needed could be unnecessarily harmful to the environment or to humans, livestock and wildlife in the proximity of the application (Dailey *et al*., 1998; Flickinger *et al*., 1991). Using less than is needed won't accomplish the control that was desired and might result in the evolution of resistance to the insecticide (Shufran *et al.,* 1996; Rider *et al.,* 1998), and insecticides may reduce natural enemy populations, thereby exacerbating problems of control (Basedow *et al*., 1985; Matacham and Hawkes, 1985; Croft, 1990). Generally, that dose is referred to as an effective dose-95 or ED_{95}. Two other analogous applications

Table 4. Simulation results using two confidence intervals for the mean from the same normal population.

Sample size	95% CI overlap	S.E int. overlap	84% CI overlap
5	0.995	0.787	0.953
10	0.995	0.815	0.949
15	0.995	0.825	0.951
20	0.995	0.836	0.953
25	0.995	0.84	0.955
30	0.994	0.836	0.951
35	0.993	0.833	0.951
40	0.995	0.837	0.954
45	0.994	0.837	0.954
50	0.995	0.838	0.952

Each row presents the results of 10,000 pairs of simulated data sets. "Overlap" columns represent the probability the intervals overlap.

for such an analysis are the derivations of ED_{50}s for insect pathogens (e.g., Kariuki and McIntosh, 1999) and of detectability half-lives for prey proteins or DNA sequences in predator gut analysis (Greenstone and Hunt, 1993; Chen *et al.,* 2000). Confidence intervals, often referred to as fiducial limits or inverse confidence limits, can be calculated on effective dosages.

For insecticide trials, the ED is often called the lethal dose (LD). The probability of killing an insect given a specific dose is often estimated with probit regression (Ahmad et al. 2003; Smirle et al. 2003). If there are two or more independent groups of insects, it may be of interest to estimate, say, the LD_{90} for each with probit regression for the purpose of deciding which are the same. One way to do this was provided by Robertson and Preisler (1992) which involved calculating a confidence interval for the ratio of LDs. The resulting confidence interval can then be used to test the equality of the two LDs (i.e., if the value 1 is contained in the interval for the ratio, then the LDs are not significantly different). This procedure, though not difficult to perform, is not available in standardized statistical software packages such as SAS. Thus researchers might be tempted to check the overlap of fiducial limits as a substitute for the procedure outlined in Robertson and Preisler. The problem exists in this situation as it does in the case to test two means from a normal distribution. If the researcher uses 95% fiducial limits, then checking whether they overlap will result in a very conservative test. What we wish to investigate here is whether fiducial limits for each population's LD_{90} can be calculated in a way that will allow us to determine whether the values are significantly different by whether or not the intervals resulting from these fiducial limits overlap.

Ironically, Robertson and Preisler (1992) suggest this very idea. They write "Many investigators have used a crude method to address this question. They compare lethal doses by examining their 95% confidence limits. If the limits overlap, then the lethal doses do not differ significantly except under unusual circumstances." They continue with an example using a fictitious scientist named Dr. Maven. Dr. Maven wanted to compare the LD_{90} for a parent generation to that of a second laboratory generation. Robertson and Preisler continue: "The 95% confidence limits of these LD_{90}s do not overlap, and Dr. Maven concludes that they probably differ significantly. However, the exact significance level for this procedure is not clear: it is not 5%."

The fiducial limits that can be calculated on each effective dose can be used to perform the desired test. Suppose fiducial limits of some predetermined size (say $(1-\alpha)100\%$) were calculated for each population. If the fiducial limits overlapped, then the two effective dosages would be declared not significantly different. If the limits did not overlap, then the effective dosages would be declared significantly different. The primary issue at hand is to determine what α is needed for these limits to assure that the desired error levels for testing the LDs are attained. Up to now, SAS was unable to provide anything other than 95% inverse confidence limits. Beginning with SAS Version 8.2, an option was made available for the MODEL statement in PROC PROBIT that allows the user to calculate any fiducial limit he or she desires. This can be achieved by placing a /ALPHA = value after the model statement, where "value" is the decimal alpha level desired for the fiducial limit.

In order to assess the effectiveness of this proposed procedure, we performed a simulation study in PC SAS Version 8.2. The first objective is to find an appropriate level to set the fiducial limits so that they give a 0.05 test. This was accomplished by generating 5000 pairs of independent sets of binary data, with equal sample sizes of 40, from the same population (probit intercept = 0 and slope = 1). For each set of data, effective doses were calculated for the 50th, 75th, 90th, and 99th levels of probability. Fiducial limits were calculated using alpha values ranging from 0.05 to 0.20 and the number out of the 1000 pairs that overlapped

Table 5. Simulation results using two inverse confidence intervals from probit regressions performed on the same population.

Fiducial limit	Overlap for LD_{50}	Overlap for LD_{75}	Overlap for LD_{90}	Overlap for LD_{99}
0.05	0.994	0.997	0.998	0.998
0.06	0.992	0.996	0.997	0.998
0.07	0.989	0.993	0.995	0.996
0.08	0.987	0.99	0.993	0.994
0.09	0.984	0.987	0.99	0.992
0.1	0.978	0.984	0.985	0.988
0.11	0.975	0.978	0.979	0.984
0.12	0.97	0.975	0.976	0.979
0.13	0.964	0.97	0.972	0.973
0.14	0.958	0.963	0.968	0.968
0.15	0.952	0.959	0.963	0.962
0.16	0.947	0.953	0.957	0.956
0.17	0.94	0.947	0.95	0.952
0.18	0.933	0.941	0.943	0.946
0.19	0.926	0.933	0.935	0.938
0.2	0.921	0.928	0.926	0.929

These are the results of 5,000 pairs of simulated data sets. "Overlap" columns represent the probability the fiducial intervals for that particular LD level overlap.

was noted. Robertson and Preisler's method was also performed for each pair to investigate how it performed. Table 5 presents the simulation results for the proposed method. Note that a 0.95 probability of overlap occurs generally around α = 0.15-0.17, depending upon which effective dose is being tested. This is consistent with the findings in the first section of this paper in which 83% or 84% confidence intervals were found to work well in the comparison of normal means. Table 6 presents the results of Robertson and Preisler's ratio method for comparing LDs. One should note that, at least from this simulation, their method tends to reject too frequently when comparing LD_{50}s, but seems to work well at the other LDs exhibited.

An analysis of the powers of the proposed method using an adjusted fiducial alpha of 0.17 as compared to the ratio method presented in Robertson and Preisler is presented in Table 7. Different ratios of slopes of two models were generated, and the probability of rejecting the hypothesis that the LDs were the same calculated for each method. This was done for tests for LD_{50}, LD_{90} and LD_{99}. As can be seen in Table 7, the method of comparing fiducial limits is not as powerful as the ratio method. As the differences in slopes of the two probit regressions get larger (and hence, the differences in LDs), the ratio method becomes more likely to detect these differences relative to the method of comparing fiducial limits.

Table 6. Simulation results using the ratio method to test LDs (Robertson & Preisler, 1992).

Probability of Rejection for LD_{50}	Probability of Rejection for LD_{75}	Probability of Rejection for LD_{90}	Probability of Rejection for LD_{99}
0.08	0.058	0.054	0.045

These are the results of 5,000 pairs of simulated data sets.

Conclusions

Caution should be exercised when the results of an experiment are displayed with confidence or standard error intervals. Whether or not these intervals overlap does not imply the statistical significance of the parameters of interest. If the researcher wishes to use confidence intervals to test hypotheses, it appears that when the standard errors are approximately equal, using 83% or 84% size for the intervals will give an approximate $\alpha = 0.05$ test. Theoretical results for large samples as well as simulation results for a variety of sample sizes show that using 95% confidence intervals will give very conservative results, while using standard error intervals will give a test with high type I error rates. When applying this idea to test lethal doses or effective doses for two independent probit regressions, with the two populations being the same under the null hypothesis and the sample sizes being equal, using 83% level for fiducial limits will approximate a 0.05 test. However, the ratio test provided in Robertson and Preisler (1992) should be used to test effective doses since it has been demonstrated to be a more powerful

Table 7. Simulation results comparing powers of ratio test to use of fiducial limits to test differences in LD 50s, LD 90s and LD 99s in probit regressions.

Ratio of slopes	Ratio test for LD_{50} rejection rates	83% CI failure to overlap for LD_{50}	Ratio test for LD_{90} rejection rates	83% CI failure to overlap for LD_{90}	Ratio test for LD_{99} rejection rates	83% CI failure to overlap for LD_{99}
1.25	0.066	0.049	0.066	0.066	0.069	0.064
1.5	0.076	0.062	0.136	0.13	0.166	0.156
1.75	0.095	0.086	0.191	0.179	0.246	0.222
2	0.091	0.085	0.273	0.226	0.366	0.294
2.25	0.118	0.101	0.341	0.274	0.442	0.348
2.5	0.144	0.131	0.41	0.337	0.533	0.407
2.75	0.171	0.162	0.48	0.37	0.603	0.451
3	0.197	0.184	0.531	0.391	0.664	0.463
3.25	0.221	0.211	0.571	0.424	0.72	0.509
3.5	0.285	0.276	0.621	0.461	0.757	0.545
3.75	0.314	0.308	0.697	0.525	0.813	0.606
4	0.322	0.308	0.683	0.51	0.81	0.579
4.25	0.372	0.352	0.715	0.551	0.836	0.608
4.5	0.38	0.369	0.742	0.531	0.863	0.586
4.75	0.415	0.399	0.775	0.579	0.897	0.636
5	0.455	0.439	0.807	0.61	0.902	0.66

These are the results of 1,000 pairs of simulated data sets. An adjustment of alpha=0.17 was used in the fiducial limit procedure. Error rates of 0.05 were used for the ratio test. Ratio column refers to the ratio of one probit regression slope to the other probit regression slope. The intercepts of the two regressions are held constant. Large slope ratios reflect large differences in LDs.

method of comparison.

Acknowledgements

We thank Kris Giles (Oklahoma State University) and Jim Throne (USDA-ARS) for reviews of the manuscript.

References

Ahmad M, Arif MI, Denholm I. 2003. High resistance of field populations of the cotton aphid Aphis gossypii Glover (Homoptera: Aphididae) to pyretthroid insecticides in Pakistan. *Journal of Economic Entomology* 96: 875-878.

Basedow TH, Rzehak H, Voss K. 1985. Studies on the effect of deltamethrin on the numbers of epigeal predatory arthropods. *Pesticide Science* 16: 325-332.

Browne, RH. 1979. On visual assessment of the significance of a mean difference. *Biometrics*, 35: 657-665.

Chen Y, Giles KL, Payton ME, Greenstone MH. 2000. Identifying key cereal aphid predators by molecular gut analysis. *Molecular Ecology* 9:1887-1898.

Croft, BA 1990. *Arthropod Biological Control Agents and Pesticides*. John Wiley and Sons.

Dailey G, Dasgupta P, Bolin B, Crosson P, du Guerney J, Ehrlich P, Folke C, Jansson AM, Kautsky N, Kinzig A, Levin S, Mäler K-G, Pinstrup-Anderson P, Siniscalco D, Walker B. 1998. Food production, population growth, and the environment. *Science* 281: 1291-1292.

Flickinger EL, Juenger G, Roffe TJ, Smith MR, Irwin RJ. 1991. Poisoning of Canada geese in Texas by parathion sprayed for control of Russian wheat aphid. *Journal of Wildlife Disease* 27: 265-268.

Goldstein H, Healy MJR. 1995. The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A* 158: 175-177.

Greenstone MH, Hunt JH. 1993. Determination of prey antigen half-life in *Polistes metricus* using a monoclonal antibody-based immunodot assay. *Entomologia Experimentalis et Applicata* 68:1-7.

Gupta RC, Ma S. 1996. Testing the equality of the coefficient of variation in k normal populations. *Communications in Statistics*. 25: 115-132.

Kariuki C, McIntosh AH. 1999. Infectivity studies of a new baculovirus isolate for the control of diamondback moth (Lepidoptera:Plutellidae). *Journal of Economic Entomology* 92: 1093-1098.

Matacham EJ, Hawkes C. 1985. Field assessment of the effects of deltamethrin on polyphagous predators in winter wheat. *Pesticide Science* 16:317-320.

Payton, ME 1996. Confidence intervals for the coefficient of variation. *Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture*. 8: 82-87.

Payton ME, Miller AE, Raun WR. 2000. Testing statistical hypotheses using standard error bars and confidence intervals. *Communications in Soil Science and Plant Analysis*. 31: 547-552.

Rider SD, Dobesh-Beckman SM, Wilde GE. 1998. Genetics of esterase mediated insecticide resistance in the aphid *Schizaphis graminum. Heredity* 81: 14-19.

Robertson JL, Preisler HK. 1992. *Pesticide Bioassays with Arthropods*. CRC Press.

SAS Institute Inc. 1999. *SAS/STAT User's Guide, Version 8, 4th Edition*, SAS Institute.

Schenker N, Gentleman JF. 2001. On judging the significance of differences by examining overlap between confidence intervals. *The American Statistician*. 55: 182-186.

Shufran RA, Wilde GE, Sloderbeck PE. 1996. Description of three isozyme polymorphisms associated with insecticide resistance in greenbug (Homoptera: Aphididae) populations. *Journal of Economic Entomology* 89: 46-50.

Smirle MJ, Lowery DT, Zurowski CL. 2003. Susceptibility of leafrollers (Lepidoptera: Tortricidae) from organic and conventional orchards to azinphosmethyl, Spinosa, and *Bacillus thuringiensis*. *Journal of Economic Entomology* 96: 879-874.

Vangel, MG 1996. Confidence intervals for a normal coefficient of variation. *The American Statistician.* 50: 21-26.

Wilson, CA, Payton ME. 2002. Modelling the coefficient of variation in factorial experiments. *Communications in Statistics-Theory and Methods*. 31: 463-476.