Statistical power to detect main and interactive effects on the attributes of small-mammal populations

Eric M. Schauber and W. Daniel Edge

Abstract: Statistical power is an important consideration in the design of experiments, because resources invested in an experiment may be wasted if it is unlikely to produce statistically significant results when real effects or differences exist. Using data from toxicological experiments on seminatural populations of small mammals, we examined the power of statistical tests for main and interactive effects. Our objectives were to evaluate the efficacy of actively reducing within-treatment variation in order to increase power and compare the power provided by several response variables commonly measured in population studies. Controlling population size (N) before treatment increased power to detect effects on N but decreased power to detect effects on population growth (r). For a specified reduction in N, r provided higher power than N. Fractional measures of recruitment generally provided low power, especially when N was low (<20 animals). Power to detect an interaction of two adverse treatments depended on the magnitudes of their main effects, as well as the magnitude of interactive effects. Estimating or predicting effect size is more complex and difficult for interactive effects than for main effects. We conclude that researchers can increase the probability of detecting real effects by choosing response variables with relatively low inherent variability. However, efforts to actively reduce within-treatment variation may have unanticipated repercussions in natural systems.

Résumé : La puissance statistique est un facteur de considération important lors de la mise au point d'expériences parce que les ressources investies dans une expérience peuvent l'être en pure perte si elles ne peuvent donner lieu à des résultats statistiquement significatifs même lorsque les effets ou différences sont réels. Nous avons utilisé les résultats d'expériences toxicologiques sur des populations seminaturelles de petits mammifères pour examiner la puissance des tests statistiques dans l'étude des effets principaux et des effets d'interactions. Notre objectif était d'évaluer l'efficacité de la réduction contrôlée de la variation intra-groupe (traitement) pour augmenter la puissance et de comparer la puissance obtenue d'après plusieurs variables mesurées de façon courante dans les études démographiques. En réduisant l'effet de la taille de la population (N) avant le traitement, on obtient une augmentation de la puissance dans la détection des effets sur N, mais une diminution de la puissance dans la détection des effets sur la croissance de la population (r). Pour une réduction spécifique de N, r fournit une puissance supérieure à celle de N. Des mesures fractionnelles du recrutement résultent généralement en une puissance faible, surtout lorsque N est faible (<20 animaux). L'efficacité à détecter une interaction de deux traitements adverses dépend de l'amplitude de leurs effets principaux et aussi de l'amplitude des effets de leurs interactions. L'estimation ou la prédiction de l'amplitude des effets est plus complexe et plus difficile dans le cas des effets des interactions que dans le cas des effets principaux directs. Nous concluons que les chercheurs peuvent augmenter leurs chances de détecter des effets réels en choisissant des variables à variabilité inhérente faible. Cependant, les efforts consacrés à la réduction de la variation intra-groupe peuvent avoir des répercussions non anticipées dans les systèmes naturels.

[Traduit par la Rédaction]

Introduction

Data from ecological experiments are most often analyzed under a hypothesis-testing paradigm. A null hypothesis (H_0; e.g., that no difference exists) is posited and statistical tests are used to determine if the data are sufficient to reject this hypothesis at a specified level of significance (α), thereby supporting the alternative hypothesis (e.g., that a difference exists). Often, however, data collected during ecological research are insufficient to reject H_0 at the prescribed significance level, even though the alternative hypothesis may actually be correct. Researchers can reduce the probability of this result by designing studies in such a way as to maximize statistical power, the probability of correctly rejecting a false H_0 (Sokal and Rohlf 1969). Power is $1 - \beta$, where β is the probability of failing to reject a false H_0 (Type II error).

The power of a statistical test increases when sample size (n), effect size, or α is increased or when variation within samples or treatment groups is reduced (Cohen 1988). Effect size is the degree of departure from H_0, and may refer to a difference among groups, an odds ratio, or a regression coefficient. Sample sizes are often limited by constraints on

Received April 20, 1998. Accepted August 21, 1998.

E.M. Schauber[1] and W.D. Edge. Department of Fisheries and Wildlife, Oregon State University, 104 Nash Hall, Corvallis, OR 97331-3803, U.S.A.

[1]Author to whom correspondence should be sent at the following address: Department of Ecology and Evolutionary Biology, University of Connecticut, U-42, Storrs, CT 06269, U.S.A. (e-mail: ems95002@uconnvm.uconn.edu).

 © 1999 NRC Canada

time, effort, or materials. However, power can be maximized, despite a limited *n*, by reducing within-treatment variation. This can be accomplished by reducing measurement error, choosing response variables with relatively low inherent variability, blocking on known sources of variation, accounting for covariates, or actively reducing the variation within each treatment.

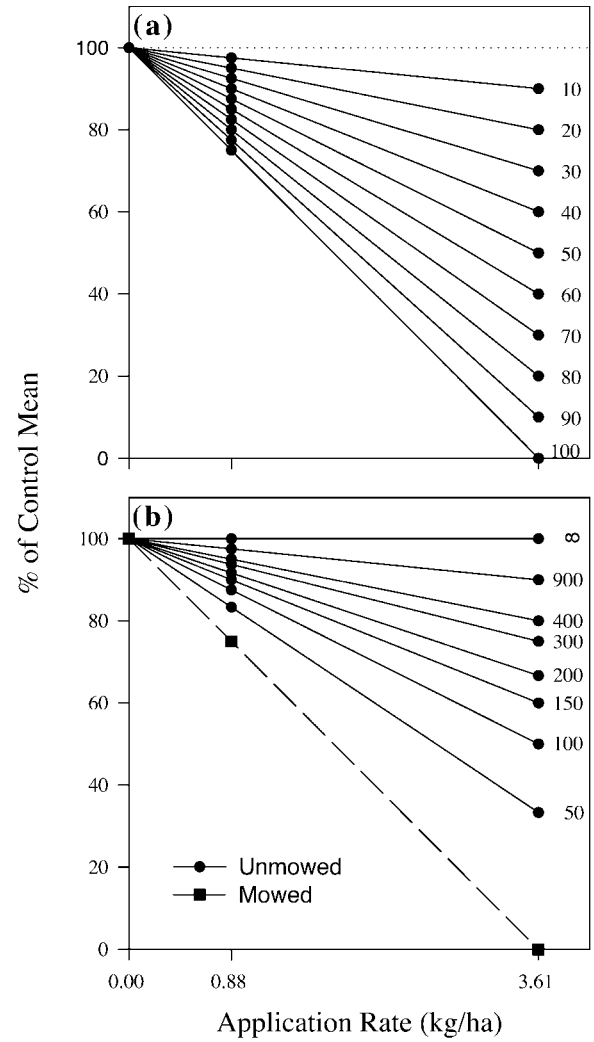
We examined the power of statistical tests for main and interactive effects of insecticide and habitat alteration on small mammals. These treatments were applied in two controlled experiments (Edge et al. 1996; Schauber et al. 1997). Our experiments were originally designed to field test assumptions of the United States Environmental Protection Agency risk-assessment methods regarding exposure of nontarget organisms to pesticides. Furthermore, in a previous experiment, we found that gray-tailed vole (*Microtus canicaudus*) populations responded to mowing in a similar manner regardless of initial vole densities (Edge et al. 1995). Our objectives for this paper were to evaluate the efficacy of actively reducing within-treatment variation and to compare the power provided by several response variables commonly measured in population studies. This information can help researchers maximize power achieved by similar studies in the future.

Materials and methods

Our experiments were conducted at the Hyslop Agronomy Farm, 10 km north of Corvallis, Oregon. This site featured twenty-four 0.2-ha enclosures designed to contain closed populations of small mammals in a habitat of alfalfa (*Medicago sativa*). In 1992, we performed an experiment to determine the effects of an organophosphorous insecticide, azinphos-methyl, applied at 0, 0.77, 1.55, 3.11, and 4.67 kg/ha, on gray-tailed voles (Edge et al. 1996). In 1993, we used a 3×2 factorial design to test whether the effects of azinphos-methyl (applied at 0, 0.88, and 3.61 kg/ha) on gray-tailed voles, deer mice (*Peromyscus maniculatus*), and house mice (*Mus musculus*) differed between mowed (alfalfa height reduced by 70%) and unmowed enclosures (Schauber et al. 1997). In both experiments, four replicate enclosures were assigned to each treatment or combination of treatments. For all species of test organism, each enclosure was treated as a single experimental unit. We hypothesized that both mowing and insecticide application (main effects) would adversely affect the small mammals. We also hypothesized that effects of the insecticide in 1993 would be more severe in enclosures where alfalfa had been mowed (i.e., treatment effects would interact) because more of the chemical would reach ground level.

We used mark–recapture methods, detailed in Edge et al. (1996) and Schauber et al. (1997), to estimate the population size (*N*) and growth rate (*r*), the proportion of adult females in reproductive condition (pregnant or lactating), the proportion of each population composed of recruits, and the number of young recruits per adult female. Because of their small populations, we did not measure reproductive parameters for house mice. Population growth rate was calculated by $r = [\ln (N_t/N_{t-x})]/x$, where x is the number of weeks between 4-night trap periods. Vole populations in insecticide-free enclosures were ≥20 individuals per enclosure for most of both experiments, and exceeded 100 in some enclosures. Deer mouse and house mouse populations were much smaller, rarely exceeding 15 individuals of either species per enclosure. In 1993, to reduce within-treatment variation in *N*, we kept vole populations at approximately 15 adults of each sex per enclosure by removing young and untagged voles each trap period until 2 weeks before the insecticide was applied. All experimental animals were cared for in accordance with the principles and guidelines of the Canadian Council on Animal Care, using an approved animal welfare protocol.

Fig. 1. Specified hypothetical effect sizes of azinphos-methyl application (percent reduction (*a*)) and interaction between azinphos-methyl application and mowing treatment (percent greater effect in mowed enclosures (*b*)) on measured response variables used in power analysis.



After using multivariate repeated-measures analysis of variance (ANOVA) to determine that time × treatment interactions were significant, we used univariate ANOVA to test for differences among treatments during each trap period in the studies. To increase statistical power we used $\alpha = 0.1$. We examined residual plots to check for violations of ANOVA assumptions. We analyzed ln (*N*) of vole and ln (*N* + 1) of house mouse populations and ln (x + 0.05) of recruits per adult female to correct skewness and stabilize variances. The constants we added for these transformations were the lowest observed positive values for these measures. *N* for deer mice met the assumptions of ANOVA without transformation. We used an arcsine square root transformation to stabilize variances of the proportion of females in reproductive condition and the proportion of each population composed of recruits.

We estimated the power of ANOVA tests for main effects of insecticide application and interactive effects of mowing and insecticide treatments on each response variable for each species, for the

© 1999 NRC Canada

Fig. 2. Interexperimental differences in power to detect specified insecticide effects on size (*a*) and growth rate (*b*) of gray-tailed vole populations. Symbols and error bars denote means ± 1 SD for four trap periods after pesticide application.

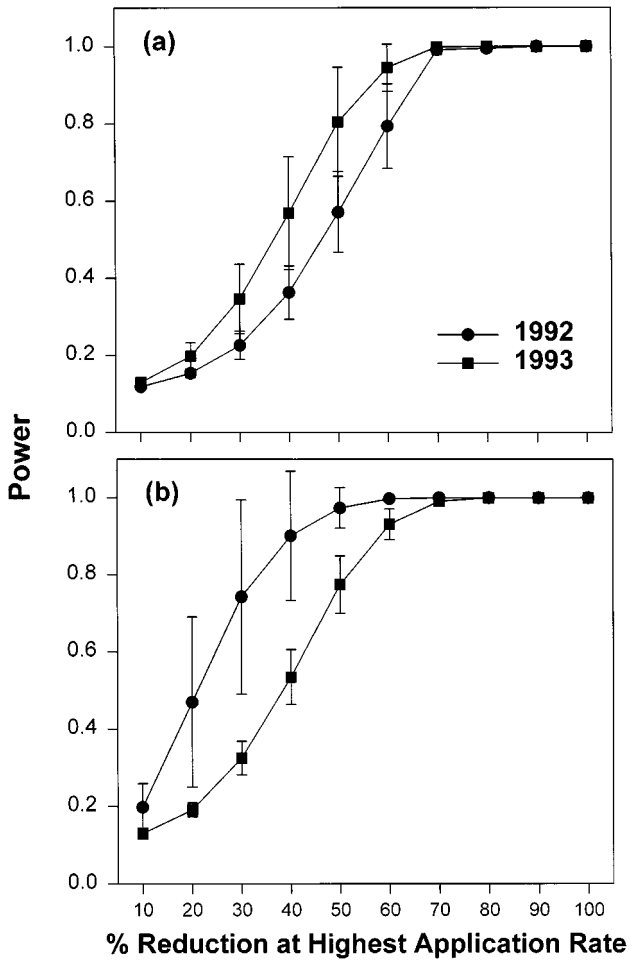
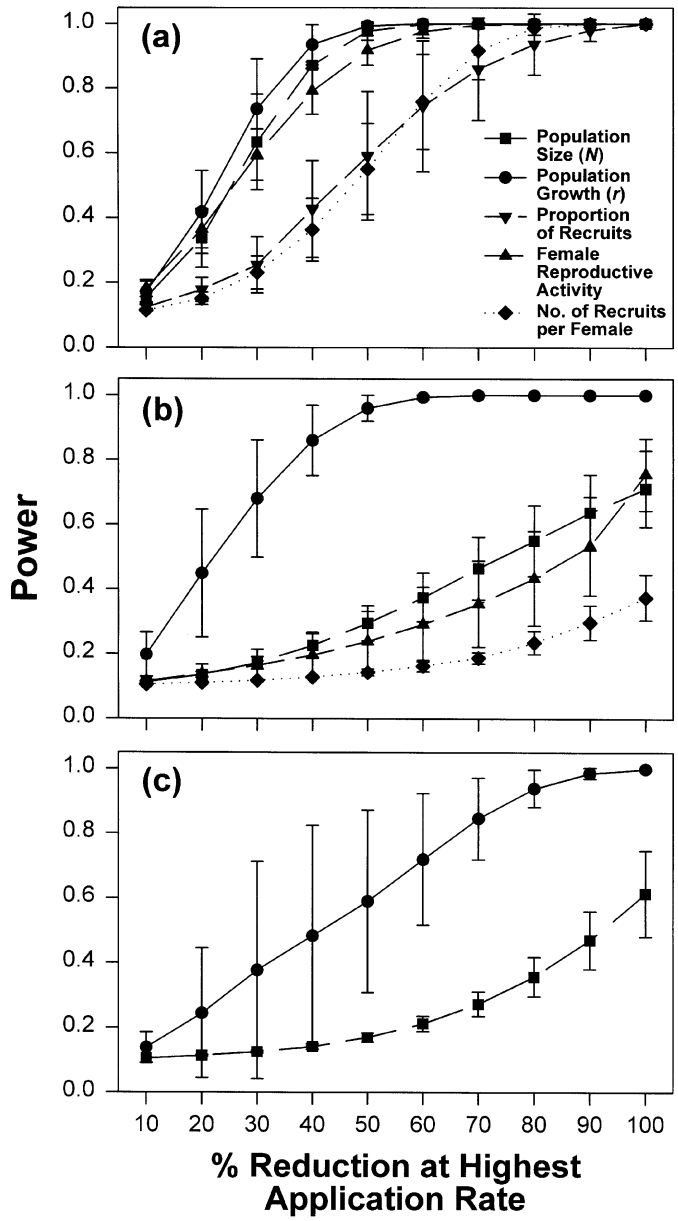


Fig. 3. Power to detect specified effects of azinphos-methyl on demographic parameters measured in 1993 for gray-tailed voles (*a*), deer mice (*b*), and house mice (*c*). Symbols and error bars denote means ± 1 SD for four trap periods after pesticide application (two trap periods for deer mouse recruitment).



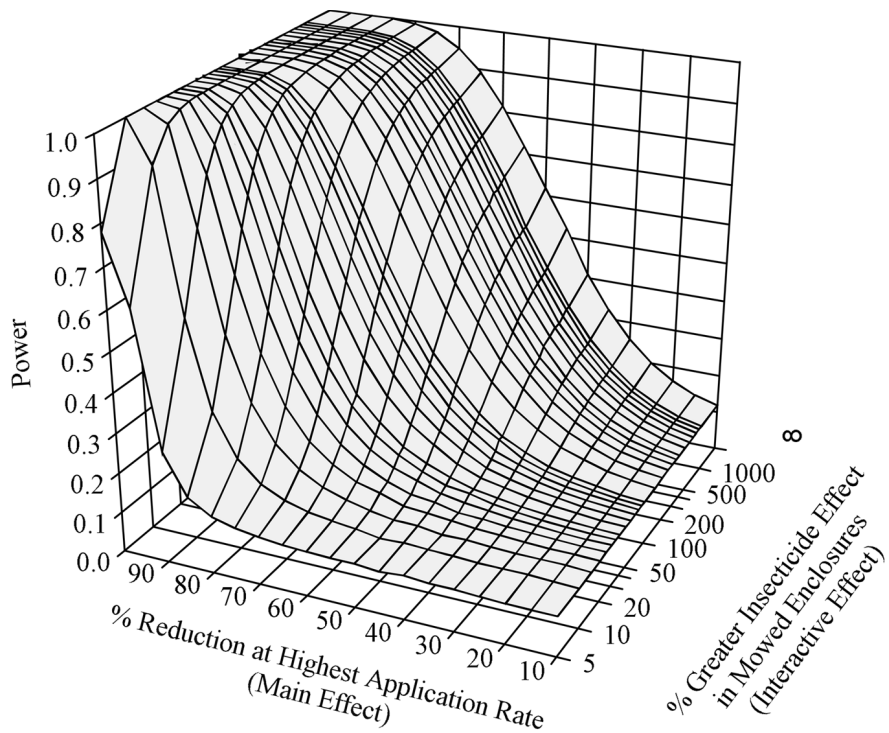
first four trap periods or trap-period intervals after insecticide application. We captured deer mouse recruits in mowed enclosures for only two trap periods after insecticide application, so we present the power for recruits per female for those trap periods only and not the proportion of recruits for deer mice. The noncentrality parameter, ϕ, and degrees of freedom were calculated for each test for insecticide effects and insecticide × mowing interactions (Neter and Wasserman 1974), and the power for $\alpha = 0.1$ was determined using published power values (Tiku 1972). Pooled sample standard deviation, used to calculate ϕ, was calculated for each response variable and sampling occasion by taking the square root of the mean square error from the ANOVA table.

We estimated the power for a range of effect sizes. Effects of the insecticide were specified as 10–100% reductions from control (0 kg/ha) means at the highest application rate, assuming that effect size increased linearly with application rate (Fig. 1*a*). Sample means used to calculate ϕ were determined by applying these hypothesized effects to control values, transforming if appropriate, and averaging. For tests on *r*, we determined treatment effects by applying specified reductions in *N* between consecutive trap periods. For example, a 50% reduction in *N* over 2 weeks produced $r = [\ln(0.5)]/2 = -0.347$, which was added to the mean *r* for control populations. Thus, specified effects on *N* and *r* were similar, in that both referred to reductions in population size. This allowed direct comparisons of the power provided by the two responses.

Specifying effect sizes for interactive effects was much more complex than for main effects. We expected that untransformed means would decrease linearly with increasing application rate, but that this effect would be greater in mowed than in unmowed enclosures. The effect size for this interaction, therefore, was the difference in the effect of the insecticide between mowed and unmowed enclosures. For interactive effects, we established a linear insecticide effect in mowed enclosures, resulting in an untransformed mean *N* of 0 (i.e., 100% reduction) at 3.61 kg/ha and linear insecticide effects in unmowed enclosures, so that effects in the mowed enclosures were 5% greater to infinitely greater (no effect in unmowed enclosures) than in unmowed enclosures (Fig. 1*b*). To

© 1999 NRC Canada

Fig. 4. Influences of main-effect size (maximum insecticide related reduction) and interactive-effect size (difference between mowed and unmowed enclosures in insecticide-related reduction) on power to detect interactive effects of azinphos-methyl and mowing on gray-tailed vole densities.



investigate the influence of main-effect size on the power to detect interactive effects, we also calculated the power for this interaction when the main effect of the 3.61 kg/ha rate in the mowed enclosures ranged from a 5 to a 95% reduction from control-enclosure means.

To determine the efficacy of controlling within-treatment variation, we compared power to detect specified effects on *N* and *r* for vole populations during the 1993 experiment, when they were kept at about 30 voles/enclosure, with data from 1992, when populations were not controlled. Because experimental designs differed between the 2 years, we used the 1992 analytical design (one-way ANOVA with five factor levels) to compare between years. For 1993 data, we used means from unmowed, 0 kg/ha enclosures as control means, to allow comparison with data from 1992, when enclosures were not mowed. Finally, we compared power curves among the response variables measured in 1993 to determine which population measures provided the highest and lowest power.

Results

Effects of reducing within-treatment variation

At the time of insecticide application, *N* for voles ranged from 16 to 106 per enclosure in 1992, when *N* was not controlled (Edge et al. 1996), but only from 14 to 61 per enclosure in 1993, when *N* was controlled (Schauber et al. 1997). Power to detect specified effect sizes on ln-transformed *N* for voles was greater for data collected in 1993 than for data collected in 1992 (Fig. 2*a*). For example, to detect an insecticide effect resulting in a 50% reduction in *N* (before transformation) in enclosures treated at the highest application rate, power was 0.81 for 1993 data but only 0.57 for 1992 data. However, power to detect effects on *r* was lower for data collected in 1993 than for data collected in 1992 (Fig. 2*b*). By analyzing *r*, power to detect a 50% reduction in *N* between two trap periods for the high-rate treatment was 0.97 in 1992 but only 0.78 in 1993.

Differences in power among response variables

For deer mice and house mice, which had small populations ($N < 20$/enclosure), greater power to detect specified effects on *N* was achieved by measuring *r* than by measuring *N* (Figs. 3*b* and 3*c*). For voles in 1993, power curves for *N* and *r* were similar (Fig. 3*a*). For voles and deer mice, power for specified effects varied among measurements of reproductive parameters (Figs. 3*a* and 3*b*). The proportion of adult females in reproductive condition provided higher power for both species than did the proportion of captured animals composed of recruits. The number of recruits per adult female provided the lowest power, because of high within-treatment variation (Figs. 3*a* and 3*b*).

Power to detect interactive effects

Power to detect interactive effects of insecticide and mowing was affected by both the magnitude of insecticide effects in both mowing treatments and the difference in insecticide effects between mowing treatments (Fig. 4). For example, if the effect of the insecticide on *N* for voles during the first trap period after application was twice as great in mowed as in unmowed enclosures, then the power to detect this interaction would have been 0.96 if the application rate of 3.61 kg/ha caused a 40 and 80% reduction in unmowed and mowed enclosures, respectively. However, the power to detect the interaction would have been only 0.39 if the same application rate reduced vole populations in

© 1999 NRC Canada

unmowed and mowed enclosures by 30 and 60%, respectively.

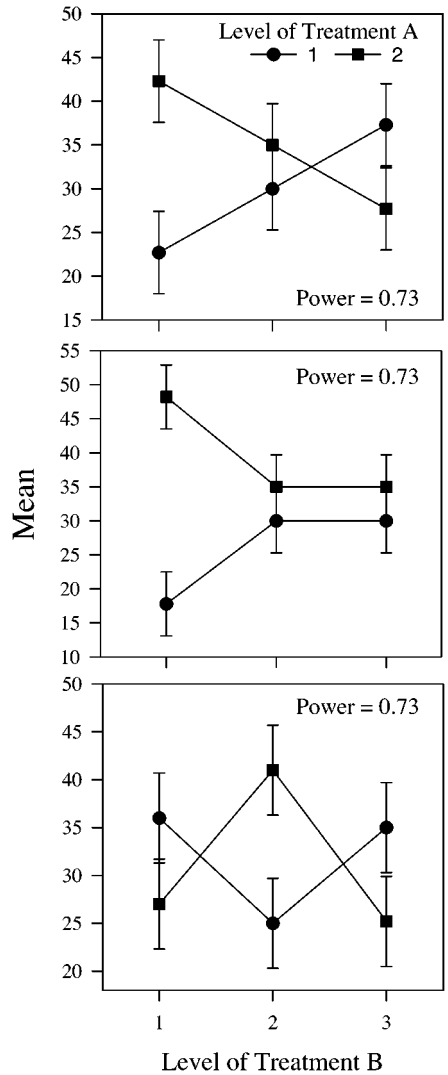
Discussion

Estimating power to detect interactive effects of mowing and insecticide proved to be complicated, and was sensitive to the magnitude of main effects as well as interactive effects. We recommend that researchers perform power analyses before conducting an experiment designed to test for interactive effects, examining a range of main- and interactive-effect sizes, because small changes in either type of effect could affect power substantially. Estimating power to detect interactive effects is also complicated by the diversity of forms that interactions can take (Fig. 5). For our experiment, the expected interaction was relatively simple: an increase in the magnitude of a linear insecticide effect. However, main effects can be nonlinear and interactive effects may not be so straightforward. Researchers hoping to estimate the power to detect interactions a priori must be very explicit about the expected form of both interactive and main effects, and should explore many alternative forms of these effects.

By maintaining vole populations near 15 pairs/enclosure, we reduced within-treatment variation and increased our power to detect insecticide effects on *N*. Evidently, however, this manipulation decreased power to detect effects on *r*. By moving animals among enclosures, we may have disrupted the reproductive cycles of transplanted voles and affected social structure, adversely affecting reproduction and *r*. Because the number of voles that we introduced differed among enclosures, this manipulation probably increased within-treatment variation in *r*, thereby reducing power. Thus, manipulating populations to decrease within-treatment variation in one demographic variable may increase variation in other variables. Density-dependent effects on *r* may occur at high population densities, resulting in subsequent loss of power if densities are not reduced. However, we have found few density-dependent effects in our previous studies (Edge et al. 1995, 1996; Schauber et al. 1997), even though vole densities within our experimental enclosures are much higher than densities of natural vole populations (Wolff et al. 1996).

Power to detect insecticide effects varied among response variables. When densities were not controlled, as was the case for voles in 1992 and deer mice and house mice in 1993, *r* provided greater power to detect a specified effect on small-mammal populations than did *N*. This suggests that *r* for uncontrolled small-mammal populations may be inherently less variable than *N*. *N* is a product of cumulative influences of historical growth rates. However, *r* may be more responsive to immediate environmental conditions, therefore variance in *r* likely will be lower than that in *N* among populations experiencing similar environmental conditions. When population changes are rapid, as with acute exposure to insecticide, lower variance in *r* than in *N* results in greater statistical power. For detecting chronic low-intensity effects, however, *N* may provide greater power than would *r*, because small changes in *r* can result in large differences in *N* over time. Researchers testing treatments hypothesized to have acute effects on *r* probably can increase power by using simultaneous replicate populations inhabiting similar habitats in close proximity to one another.

Fig. 5. Some possible treatment interactions in a 3 × 2 factorial design, with associated power of ANOVA to detect an interaction with $\alpha = 0.05$, $n = 4$ per treatment combination, and pooled SD = 9.4. Symbols and error bars denote means ±1 SE.



For most toxicological studies, population size and growth and survival rates are the primary parameters of interest. However, impairment in reproduction and recruitment may also be of interest. The proportion of captured animals composed of recruits and the number of recruits per adult female provided low power, especially for deer mice, which were much less abundant than voles. For both measures of deer mouse recruitment, power was <0.6, even for a 100% reduction in enclosures treated with insecticide at 3.61 kg/ha. A small *N* value increased variation in the number of recruits entering the population during a trap period, as well as the total number of animals and the number of adult females captured. Because the numerators and denominators of our fractional metrics were generally small integers, integral changes in either the numerator or the denominator resulted in relatively large changes in the fraction. For example, one enclosure might have three recruits and three females (= 1

© 1999 NRC Canada

recruit/female), whereas another might have three recruits and two females (= 1.5 recruits/female), a 50% difference in the number of recruits per female. Thus, fractional metrics can be expected to have a large variance and to provide relatively low power when populations are small. These problems with fractional metrics are further exacerbated in mark–recapture studies, when capture probabilities are below 1 because random variation determines which animals are captured. For example, a capture probability of 0.68 for female voles (Edge at al. 1995) in a population of 20 females, 50% of which were pregnant, could result in observed pregnancy rates ranging from 0.29 to 0.71. Increasing capture probabilities could reduce variation in proportional metrics and, subsequently, increase power. Otis et al. (1978) discuss options for increasing the probability of capture.

Statistical power has been compared among alternative response variables commonly used in toxicological experiments on aquatic microcosms (Pratt and Bowers 1992) and in monitoring contaminated aquatic systems (Osenberg et al. 1994), but responses frequently measured in terrestrial population ecology have been ignored. Further comparisons are necessary to identify biologically relevant parameters that are sensitive to perturbations, yet possess little inherent variability, and can be measured precisely. Concentrating effort on measuring such parameters can increase the probability of detecting ecologically significant responses.

Acknowledgments

P. Bayley, H. Brunkal, J. Peterson, R. Steidl, J. O. Wolff, and two reviewers provided thoughtful criticisms of the manuscript. L. Ganio's statistical assistance was extremely valuable. These experiments were funded by the United States Environmental Protection Agency, Environmental Research Laboratory, Corvallis, Oregon (Cooperative Agreement No. CR818517). This is manuscript No. 10 945 of the Oregon Agricultural Experiment Station. Publication was supported in part by the Thomas. G. Scott Achievement Fund.

References

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Academic Press, New York.

Edge, W.D., Wolff, J.O., and Carey, R.L. 1995. Density-dependent responses of gray-tailed voles to mowing. J. Wildl. Manage. **59**: 245–251.

Edge, W.D., Carey, R.L., Wolff, J.O., Ganio, L.M., and Manning, T. 1996. Effects of Guthion 2S® on *Microtus canicaudus*: a risk assessment validation. J. Appl. Ecol. **33**: 269–278.

Neter, J., and Wasserman, W. 1974. Applied linear statistical models: regression, analysis of variance, and experimental designs. Richard D. Irwin, Inc., Homewood, Ill. p. 584.

Osenberg, C.W., Schmitt, R.J., Holbrook, S.J., Abu-Saba, K.E., and Flegal, A.R. 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. Ecol. Appl. **4**: 16–30.

Otis, D.L., Burnham, K.P., White, G.C., and Anderson, D.R. 1978. Statistical inference from capture data on closed animal populations. Wildl. Monogr. No. 62. pp. 77–80.

Pratt, J.R., and Bowers, N.J. 1992. Variability of community metrics: detecting changes in structure and function. Environ. Toxicol. Chem. **11**: 451–457.

Schauber, E.M., Edge, W.D., and Wolff, J.O. 1997. Insecticide effects on small mammals: influence of vegetation structure and diet. Ecol. Appl. **7**: 136–150.

Sokal, R.R., and Rohlf, F.J. 1969. Biometry: the principles and practice of statistics in biological research. W.H. Freeman and Co., San Francisco. p. 163.

Tiku, M.L. 1972. More tables of the power of the *F*-test. J. Am. Stat. Assoc. **67**: 709–710.

Wolff, J.O., Manning, T., Meyers, S.M., and Bentley, R. 1996. Population ecology of the gray-tailed vole, *Microtus canicaudus*. Northwest Sci. **70**: 334–340.

© 1999 NRC Canada