Environ. Biosafety Res. 2 (2003) –

DOI: 10.1051/ebr:2003008

## Editorial

# Negative and positive data, statistical power, and confidence intervals

**D.A. Andow**

Department of Entomology and Center for Community Genetics, University of Minnesota, St. Paul, MN 55108, USA. E-mail: dandow@dandow.email.uma.edu

What is negative about negative data? Scientists understand **negative data** from our training in data analysis and statistics, where we use a positive concept of negative data. Negative data are data that do not enable us to reject our null hypothesis. Such data are often difficult to publish because it is not possible to prove the null hypothesis. Every active research scientist has a large drawer where these data languish. In the area of environmental biosafety, however, some scientists have begun to use "negative data" in a second, normative way. This normative concept of negative data has socio-political connotations, where "negative" data has come to connote results that GMO proponents could use to support, and GMO opponents could use to oppose the development of GMOs. This politicization of GMO biosafety research is worthy of study in its own right, but *EBR* is prepared to accept any kind of "negative" or "positive" data.

## HERE IS DESCRIBED THE STATISTICAL STANDARDS THAT WILL BE ENFORCED IN FUTURE PUBLICATIONS IN *EBR*

There are three good reasons for these requirements. First and foremost, this ensures a high level of scientific quality for papers published in *EBR*. Second, because genetic engineering is politically charged, critics and proponents must have the opportunity to evaluate independently the quality of the available research. Finally, the scientific community needs to be able to build on the published literature. A sound meta-analysis of the accumulated results of many publications requires knowledge of the sufficient statistics of each of those experiments (*e.g.*, Arnqvist and Wooster, 1995). For example, 5 non-significant results could combine into a statistically significant result *via* meta-analysis, or significant results might melt into non-significance under the weight of multiple studies. Such meta-analyses would be valuable contributions to the scientific literature.

There are two types of error in any experiment. Type I error occurs if the null hypothesis is erroneously rejected when in actuality it is true. Typically the Type I error rate is 0.05, *i.e.,* a 1 in 20 chance that the null hypothesis is mistakenly rejected. This kind of error is routinely handled by conventional scientific practices. Type II error occurs when the null hypothesis is not rejected when in actuality it should have been rejected. Negative data suffers from the possibility of Type II error. Type II error is problematic, because as scientists we have been trained to minimize Type I errors and not be as concerned with Type II errors. Type II error is measured by statistical power. An experiment with high power has a low Type II error rate and an experiment with low power has a high Type II error rate.

In risk related problems, however, Type II errors can be more serious than Type I errors (*e.g.*, Hill and Sendashonga, 2002). For example to answer the question, what amount of GM-maize can be introduced without harming a non-target species, a relevant null hypothesis is that a certain quantity of *Bt* maize does not adversely affect non-target species. If an experiment with low statistical power were conducted, the probability of rejecting the null hypothesis will be low, whether or not the true effect had been biologically significant (Marvier, 2002). If *Bt* maize were introduced at that quantity, this kind of Type II error would result in adverse non-target effects when none had been expected. Thus, for risk related problems, Type II error must be considered explicitly.

## RETROSPECTIVE POWER ANALYSIS

A common, but flawed approach to dealing with Type II error is to require calculation of statistical power from the

experimental data. This analysis is called retrospective power analysis, and contrasts with prospective power analysis, which uses power calculations to make decisions about future experimental designs. The problem with retrospective power analysis is that power and Type II error are not independent of the Type I error rate chosen by the investigator (Hoenig and Heisley, 2001). Indeed, in many cases Type II error rate is an increasing function of the Type I error rate. In other words, for any particular experiment, the choice of the Type I error rate (usually 0.05) uniquely determines the retrospective power of the experiment. What is desired is a Type II error rate that is independent of the Type I error rate. With retrospective power analysis, this is not possible. Consequently, despite the fact that many statistical packages routinely provide retrospective power estimates, they will not be accepted in *EBR*.

## PROSPECTIVE POWER ANALYSIS

One acceptable approach for using estimated power is prospective power analysis. Here estimated power is used to design future experiments. This can give a useful indication of whether further testing of the hypothesis is feasible, *i.e.* without requiring analysis of an unreasonable number of samples (for example, see Bourguet et al., 2001). The power of a simple one-way fixed effects ANOVA is readily calculated from the noncentrality parameter and the noncentral F-distribution (there is a short, readable account in Oehlert, 2000). A nice example using power analysis in research planning is provided by Steidl et al., 1997). Unfortunately, the opportunity to design an experiment based on such power analysis is uncommon, and although prospective power analysis is acceptable, it will not be emphasized in *EBR*.

## CONFIDENCE INTERVALS REQUIRED

Confidence intervals are an acceptable way to indicate Type II error. Once confidence intervals are specified, power calculations provide no additional statistical insights from the experimental data (Hoenig and Heisey, 2001). For example, it is pointless to calculate power for hypotheses outside the confidence interval, because the interval already indicates that they are unlikely. Similarly, it is pointless to calculate power for hypotheses inside the confidence intervals, because the interval already indicates that these are not refuted by the data.

The crucial details are in the calculations of the mean and the confidence interval. The appropriate measurement scale must be specified, the sample units (number of independent replicates) clarified, and the appropriate mean and error variance calculated.

### Measurement scale

Transformation of the dependent variable is usually treated as an arbitrary choice made to meet the assumptions of the statistical model. For example, data are usually transformed to near-normality prior to conducting an ANOVA. For risk related problems, however, data transformations also change the way risks are characterized (Box 1). In general, environmental risks are differently perceived depending on circumstance and perspective. For resistance evolution, risks might be measured as the number of years gained or lost under different management policies (the untransformed time scale). For some non-target effects, risk might be measured as the proportional change in the population size of the non-target species (a logarithmic transformation), which estimates how many times larger or smaller the effect may be. Thus, the choice of statistical transformation of the original data is also a decision about the characterization of risk – it is not merely a statistical formality to ensure that the data meet the assumptions of the analysis. Authors should carefully specify the scale of measurement of the dependent variables.

### Replication and pseudoreplication

In many experimental designs replication occurs at numerous levels. For example, in a laboratory experiment individuals may be replicated within a trial, and trials may be replicated across time. In a field experiment (*e.g.,* Box 1), samples may be replicated within plots, and plots may be replicated in locations, and locations may be replicated in space. Pseudoreplication occurs when the investigator designates the wrong level (usually one with more degrees of freedom) as the unit of experimental replication (Hurlbert, 1984; Ramirez et al., 2000). Authors should clearly designate the unit of replication and the number of replicate samples in the experiment; too often this important detail is missing.

### Mean, variance and confidence interval

Box 2 provides an example of how to calculate appropriate confidence intervals around estimated population means and for specific statistical hypotheses. Because many of the papers that will be published in *EBR* will involve a

**Table 1.** Randomized complete block (RCB) design with 4 blocks and 2 treatments, transgenic and non-transgenic, for a total of 8 experimental plots. On each plot, the response variable was measured in 10 samples. Samples could be from plants, traps, soil samples, etc. The response variable could be plant yield, plant height, herbivore biomass, natural enemy density, etc. Data were normally distributed with means $\mu_1 = 20$ and $\mu_2 = 30$ for the two treatments and equal variances, $\sigma^2 = 25$. Each plot mean, $x_{ij}$, was a random draw from one of these normal distributions. Samples were normally distributed around the plot mean with a variance of 25; standard errors of the plot means, $se_{ij}$, were calculated.

| Block, *i* | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
|---|---|---|---|---|---|---|---|---|
| Treatment, *j* | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| Sample 1 | 7.2 | 25.4 | 23.1 | 15.1 | 22.0 | 29.3 | 21.5 | 20.6 |
| Sample 2 | 15.9 | 28.5 | 32.7 | 16.0 | 27.2 | 27.0 | 19.2 | 34.2 |
| Sample 3 | 8.3 | 24.2 | 34.5 | 23.7 | 15.2 | 24.2 | 24.9 | 19.6 |
| Sample 4 | 20.4 | 36.3 | 28.9 | 19.2 | 15.4 | 16.6 | 14.0 | 27.7 |
| Sample 5 | 6.9 | 28.8 | 20.6 | 22.7 | 18.3 | 22.5 | 12.4 | 19.2 |
| Sample 6 | 17.8 | 31.7 | 39.5 | 22.2 | 15.5 | 25.4 | 17.1 | 31.8 |
| Sample 7 | 6.5 | 30.1 | 19.2 | 23.6 | 15.9 | 24.3 | 23.9 | 34.1 |
| Sample 8 | 7.5 | 22.7 | 33.2 | 16.2 | 20.4 | 27.5 | 20.3 | 29.6 |
| Sample 9 | 9.2 | 24.5 | 16.7 | 23.0 | 16.5 | 28.2 | 19.5 | 22.8 |
| Sample 10 | 8.0 | 26.4 | 27.1 | 30.3 | 25.6 | 23.8 | 26.7 | 22.0 |
| Plot Mean, $x_{ij}$ | 10.8 | 27.9 | 27.6 | 21.2 | 19.2 | 24.9 | 20.0 | 26.2 |
| $se_{ij}$ | 1.6 | 1.3 | 2.4 | 1.5 | 1.4 | 1.1 | 1.5 | 1.9 |

*Measurement scale*. The untransformed measurement scale was used to assess the effect of the transgenic crop. If the data were yield or pest pressure, the untransformed scale would be an appropriate scale because it would measure the yield increase or decrease (change in pest pressure) associated with the transgenic treatment.
*Replication*. A cursory examination of the data, might lead an analyst to conclude that there were 10 or 40 replicates. True replication, however, was the plots, and there were 4 replicates of the treatments. The 10 samples in each plot were pseudoreplication.

**Box 1.** Data, measurement scale and replication.

comparison of two treatment means ($\mu_1$ and $\mu_2$), transgenic and non-transgenic, one possible null hypothesis is that the two treatments are not different. Specifically, $H_0$: $|\mu_1 - \mu_2| = 0$ for untransformed, absolute scales, and $H_0$: $|\ln(\mu_1) - \ln(\mu_2)| = 0$ for proportional risks (for categorical data the null hypothesis might be log odds = 0). If the means are estimated by independent samples, the author should show that the variances are equal (homoscedastic), before pooling them. The 95% CI will inform a reader if the null hypothesis can be rejected at the standard $\alpha = 0.05$.

## REPORT SUFFICIENT STATISTICS

Sufficient statistics are essential to allow critical evaluation of the interpretation of the data in the paper, for reviewers, readers, and future researchers who would conduct a meta-analysis of the published literature. Definitions of sufficient statistics can be tediously technical for non-statisticians, and for those interested in this, most statistics textbooks provide such a definition. For those not so inclined, the general guideline is that enough information about the statistical analysis must be presented so that the interested reader could reconstruct the entire analysis (Box 3). For example, many authors report only *P*-values from an analysis of variance – this is not enough to reconstruct the complete ANOVA table. There are many ways to report sufficient statistics – all of the *F*-values with numerator and denominator degrees of freedom and all error mean squares and degrees of freedom would be sufficient. The complete ANOVA table (*df*, either *SS* or *MS*, *F* and *P*) would also be acceptable. In addition, the relevant means on the transformed scale should be reported. It is possible that the ANOVA may be conducted on a measurement scale different from the scale that would be appropriate for calculating confidence intervals. In this case, an author might question if the ANOVA is truly needed.

## PRECISION OF MEASUREMENTS ON REPLICATE SAMPLES

In field experiments it is not possible to measure all of the organisms in a replicate plot, so the plot is subsampled to estimate the plot mean. For example, a subsample of

Confidence intervals can be used to estimate the true population means, $\mu_1$ and $\mu_2$, or to test several different null hypotheses. Data are from Box 1.

*True population means*. The true population means were estimated from the plot means $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$, where $n = 4$. The 95% confidence interval was calculated from the standard error of this estimate, $se_j = \left(\left(\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2/(n-1)\right)/n\right)^{1/2}$, and was $\pm se_j t_{\alpha,\nu}$ where $t_{\alpha,\nu}$ was the value of the $t$ distribution at $\alpha = 0.05$ (for a 95% confidence interval $\alpha = 1 - 0.95$) and $\nu = n - 1 = 3$ is the number of degrees of freedom (Tab. 2).

**Table 2.** Estimated population means and 95% confidence intervals.

| $t$ | $\bar{x}_j$ | $se_j$ | $t_{0.05,3}$ | 95% CI |
|---|---|---|---|---|
| 1 | 19.4 | 3.4 | 3.18 | 8.5 |
| 2 | 25.0 | 1.4 | 3.18 | 20.5 |

*Hypothesis tests.* The standard null hypothesis is $H_0$: $|\mu_1 - \mu_2| = 0$, *i.e.*, the two populations have the same mean. An equivalence test is based on the null hypothesis $H_0$: $|\mu_1 - \mu_2| > \Delta$, *i.e.*, the difference between the means is greater than $\Delta$. If society agrees that any difference less than $\Delta$ is biologically insignificant, rejecting the null hypothesis implies that the effect is biologically insignificant.

*Standard null hypothesis.* The treatments were paired within blocks, so the difference between the paired means was calculated, $d_i = x_{i1} - x_{i2}$. The average of the $d_i$, $\bar{d}$, and its standard error, $se_d$, was used to calculate $t = |\bar{d}|/se_d$ to test the standard null hypothesis, $H_0$: $|\mu_1 - \mu_2| = 0$ (Tab. 3). The probability that the null hypothesis was true is 0.32, which implied that the treatments were not statistically significantly different.

**Table 3.** Test of the standard null hypothesis.

| | $H_0$: $\|\mu_1 - \mu_2\| = 0$ |
|---|---|
| $\|\bar{d}\|$ | 5.7 |
| $se_d$ | 4.8 |
| t | 1.18 |
| p | 0.32 |

*Equivalence tests*. Equivalence tests also use $\bar{d}$ and $se_d$. It is also necessary to specify a $\Delta$ independent of the data. Smaller $\Delta$ are associated with more risk averse assessments, and would be one way to implement a precautionary approach to risk assessment. For $\Delta = 25$, any difference < 25 is considered insignificant, while for $\Delta = 5$, differences must be < 5 to be considered insignificant. If treatment differences < 5 were considered insignificant by society, then $H_0$: $|\mu_1 - \mu_2| > 5$ would be the appropriate equivalence test (first row of Tab. 4). The probability that this null hypothesis was true is 0.899 (Tab. 4), which was not significant at the 0.05 level. The null hypothesis cannot be rejected, and the treatment difference is likely to be greater than the socially insignificant level of 5. If $\Delta$ were 25, which would mean that society was concerned only about differences greater than 25, the conclusion would be different. In this case, the appropriate null hypothesis would be $H_0$: $|\mu_1 - \mu_2| > 25$ (second to last row of Tab. 4), and the probability that the null hypothesis was true is 0.027. This null hypothesis can be rejected, and the treatment difference is unlikely to be significant to society. Note that the true difference in the treatments is 5, but the data only allow us to conclude that the true difference is unlikely to be greater than 25.

**Table 4.** Equivalence tests for the null hypothesis $H_0$: $|\mu_1 - \mu_2| > \Delta$, for different values of $\Delta$. For these tests, $t = (|\bar{d}| - \Delta)/se_d$. $p$ is based on a two-tailed $t$-test with 3 degrees of freedom.

| $\Delta$ | $t$ | $p$ |
|---|---|---|
| 5 | 0.137 | 0.899 |
| 10 | 0.907 | 0.431 |
| 15 | 1.951 | 0.146 |
| 20 | 2.995 | 0.058 |
| 25 | 4.039 | 0.027 |
| 30 | 5.083 | 0.015 |

*Note on rounding*. Data and statistics are rounded prior to reporting in tables. In order not to propagate rounding errors, unrounded plot means from Table 1 were used to complete the calculations in Tables 2–4.

**Box 2.** Using confidence intervals.

Using confidence intervals can provide a full statistical analysis of the data. However, it is important to provide sufficient information so that the statistical analysis can be fully and independently evaluated. There are technical definitions of sufficient statistics, but for the purposes of *EBR*, the bold parts of Tables 2 and 3 would be sufficient. Using Table 3, any equivalence test could also be constructed.

Some contributors may prefer to use ANOVA to analyze their data (Tab. 5). Sufficient statistics are in bold; in addition, treatment means must be reported. The *P*-value associated with the treatment effect was the same as the *P*-value obtained using confidence intervals for the null hypothesis $H_0$: $|\mu_1 - \mu_2| = 0$ (Tab. 3).

**Table 5.** Sufficient statistics for an ANOVA are in bold and must also include treatment means. In addition, it would normally be important to report *P*-values.

| Source | *Df* | *MS* | *F* | *P* |
|---|---|---|---|---|
| Block | 3 | 9.207 | 0.20 | 0.889 |
| Treatment | 1 | 66.015 | 1.40 | 0.322 |
| Error | 3 | 45.865 | | |

**Box 3.** Sufficient statistics.

Sampling precision is related to the variation among samples within plots, and is the standard error of the plot means ($se_{ij}$, Tab. 1). In this case, the average standard error of the plot means across all plots is 1.6. Each standard error was estimated with 10 data points (9 degrees of freedom), so the 95% confidence interval around the estimated plot means was $\pm se_j t_{\alpha,\nu}$, which with $\alpha = 0.05$ and $\nu = 9$ was ±3.6. This implied that the plot means were estimated to ±3.6 with 95% confidence. Given that $|x_1 - x_2| = 5.7$ (Tab. 3), it can be seen that the plot means might be so poorly estimated that it would be difficult to discern any difference between the treatments.

**Box 4.** Sample precision.

Justification of measurement scale.
Number of replicates.
Treatment means with CIs.
Hypothesis tests with 95% CIs .
Sufficient statistics.
Sample precision.

**Box 5.** Required statistics for EBR.

plants, some number of traps, or some number of soil cores could be used to estimate the plot mean. If few subsamples are taken, the plot mean might be estimated poorly, and this variation will make it more difficult to detect differences between treatment means. Conversely, if enough subsamples are taken, the plot means can be well estimated, and the reader will have greater confidence in the results. Authors can report this precision as individual or average standard errors on plot means (Box 4).

## DISCUSSION

*EBR* will require reporting of confidence intervals on hypothesis tests, sufficient statistics and when appropriate, estimates of precision of subsampling effort (Box 5). With these requirements, other interested parties will be able to make independent judgments about the significance of any result reported in the journal. Interested authors can extend these requirements in several directions. For example, equivalence testing is an excellent method for evaluating null results formally (Hoenig and Heisey, 2001; Box 3). If we can formulate the problem so that we conclude that an effect is negligible if it is no greater than some difference Δ, we can formulate a new null hypothesis that the effect is large enough to be significant, $H_0$: $|D| > \Delta$, where D is the estimated treatment effect. This reverses the traditional burden of proof – one must be fairly certain that a large difference does not occur.

Controversy over the use of GMOs in the environment is likely to continue for some time into the future. Part of the controversy is fueled by scientific confusion and part by contending values. Requirements for systematic reporting of results will remove some of the scientific controversy and enable the discussion of the normative issues involved in risk assessment to rely on a sound scientific analysis.

## ACKNOWLEDGEMENTS

I would like to thank Kenneth Schoenly, Jason Harmon, Jennifer White and Erin Hladelik for their comments on this editorial.

## REFERENCES

**Arnqvist G, Wooster D** (1995) Meta-analysis: Synthesizing research findings in ecology and evolution. *Trends Ecol. Evol.* **10**: 236–240

**Bourguet D, Chaufaux J, Micoud A, Delos M, Naibo B, Bombarde F, Marque G, Eychenne N, Pagliari C** (2002) *Ostrinia nubilalis* parasitism and the field abundance of non-target insects in transgenic *Bacillus thuringiensis* corn (*Zea mays*). *Environ. Biosafety Res.* **1**: 49–60

**Hill RA, Sendashonga C** (2003) General principles for risk assessment of living modified organisms: Lessons from chemical risk assessment. *Environ. Biosafety Res.* **2**: (pages will be inserted later)

**Hoenig JM, Heisley DM** (2001) The abuse of power: The pervasive fallacy of power calculations for data analysis. *Am. Stat.* **55**: 19–24

**Hurlbert SH** (1984) Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**: 187–211

**Marvier M** (2002) Improving risk assessment for nontarget safety of transgenic crops. *Ecol. Appl.* **12**, 1119–1124

**Oehlert GW** (2000) A first course in design and analysis of experiments. WH Freeman, New York

**Ramirez CC, Fuentes-Contreras E, Rodriguez LC, Niemeyer HM** (2000) Pseudoreplication and its frequency in olfactometric laboratory studies. *J. Chem. Ecol.* **26**: 1423–1431

**Steidl RJ, Hayes JP, Schauber E** (1997) Statistical power analysis in wildlife research. *J. Wildl. Manag.* **61**: 270–279