

## Putting practices and products to the test of statistics

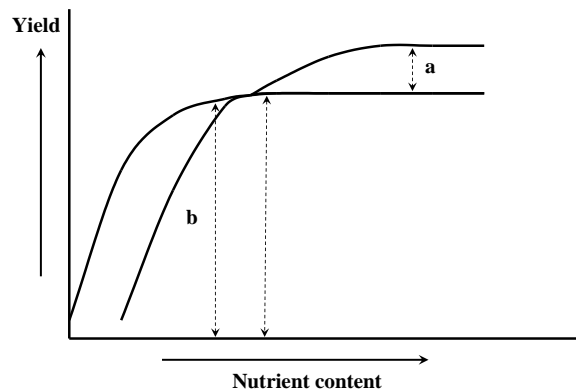
R.E. Karamanos, Viterra, 10517 Barlow Trail SE, Calgary, AB T2C 4M5

E-mail : [rigas.karamanos@viterra.ca](mailto:rigas.karamanos@viterra.ca)

These days, farmers and agronomy practitioners are inundated with new products and practices. A good marketing strategy and campaign does not necessarily imply that practices and products have been properly evaluated. Often, testimonials are being used to support a practice or a product; however, testimonials are just stories, not scientific data. People that offer testimonials are often invested in their own story and insist that what worked for them will work for everybody. Further, results from limited sites or inadequately planned experiments are utilized. Random events can result in positive effects by a product or a practice in a single trial or site that may not be reproducible. Hence proper statistical analysis must be carried out. But proper statistical analysis predicated proper experimental protocol that should be accompanied by careful and well-controlled experimental techniques must be employed with any experimental design. Normally, peer reviewed articles undergo the test of statistics and should be used as credible sources of information. Otherwise, a few rules and attention to a number of issues should be taken into consideration. This presentation aims at helping conference participants with the latter.

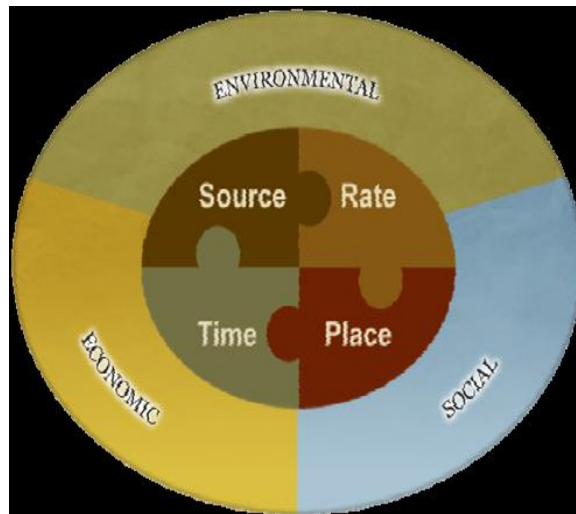
### Introduction

In a paper that I presented back in 1992 on root region management products (Karamanos, 1992), I distinguished products into nitrogen fixers and “others” and proceeded to classify the “others” into two categories: miracle products and real products. By extension, I believe all products fall into these categories. I become very apprehensive and suspicious when a product claims to provide benefit “a” in Fig. 1 compared to the “best management practices” (BMP) with conventional fertilizer products. If this were the case either the “best” agricultural practice was not being used or there was something wrong with the process of providing benefit “a”. This I believe is because there is a certain genetic potential that a crop can attain under a specific set of environmental conditions and only manipulation of the genetic potential of a crop could provide a “real” benefit of this type.



**Figure 1.** How “miracle” (a) and “real” (b) new products work.

In the same meeting I also stated that “One of the major hurdles in bringing these products to market is the fact that most of the manipulation of the yield is taking place on the upper part of the yield curve, where proportional increases or differences often become a statistical nightmare to prove, as they are obscured by high natural variability. Today with new products and practices making it into the market, adherence to the 4R nutrient management principle (Fig. 2) has become ever challenging.



**Figure 2.** The 4R nutrient stewardship concept defines the right source, rate, time, and place for fertilizer application as those producing the economic, social, and environmental outcomes desired by all stakeholders to the plant ecosystem (Bruulsema, 2009)

In a world of phenomenal media communication, marketing of products and practices has become both glamorous and wide-spread and easily accessible. However, a good marketing strategy and campaign does not necessarily imply that practices and products have been properly evaluated. The layman often neither understands nor cares whether products and practices have withstood the test of statistics. Further, change in philosophy on field research on productivity issues by both government agencies and private industry has resulted in very little research carried out with properly established protocols. Here are some common sense rules of thumb when addressing new products and practices:

- ✓ If it is too good to be true, it probably is
- ✓ Beware of hype. Hype hurts!
- ✓ Don't trust testimonials, because they are anecdotes; they are stories, not scientific data
- ✓ Look for the flipside; what did the opposing side say?
- ✓ ULTIMATE STANDARD. Is there published peer reviewed evidence?
- ✓ No substitute for thinking critically and thinking for yourself

Ultimately, the decision is for the producer, consultant, etc. to make; however, having some basic understanding of the value statistics without necessarily needing to perform them, can lead to elimination of waste of money and resources.

### **Steps in the Statistical Process**

Statistical analysis is a two-step process (Peterman 1990):

Step 1. A statistical analysis will either reject the null hypothesis ( $H_0$ ) or not. Incidentally, null hypothesis is in essence a proposal that there is no statistical significance in a set of given observations, or that a single variable is no different than zero. It is presumed to be true until statistical evidence proves otherwise.

Step 2.  $\beta$  or detectable effect size must be calculated, if  $H_0$  is not rejected when it should have been. This is what is known as statistical power. Unfortunately, decisions are made without going through the second step.

More specifically, statistical power is defined as  $1-\beta$ , where  $\beta$  is the probability of failing to reject the  $H_0$  when in fact  $H_0$  is false. In other words, the statistical power reflects the probability of correctly rejecting  $H_0$ . Ideally, the statistical power should be at least .80 to detect a reasonable departure from the null hypothesis. Table 1 summarizes the four possible outcomes for a statistical test of some null hypothesis (adapted from Toft and Shea 1983)

**Table 1.** Possible outcomes for a statistical tests and types of error

State of Nature	Decision	
	The null hypothesis is true	The null hypothesis is false
The null hypothesis is actually true	Correct ( $1-\alpha$ )	Type II error
The null hypothesis is actually false	Type I error	Correct ( $1-\beta$ )

Nelson and Rawlings (1983) described ten common misuses of statistics as follows:

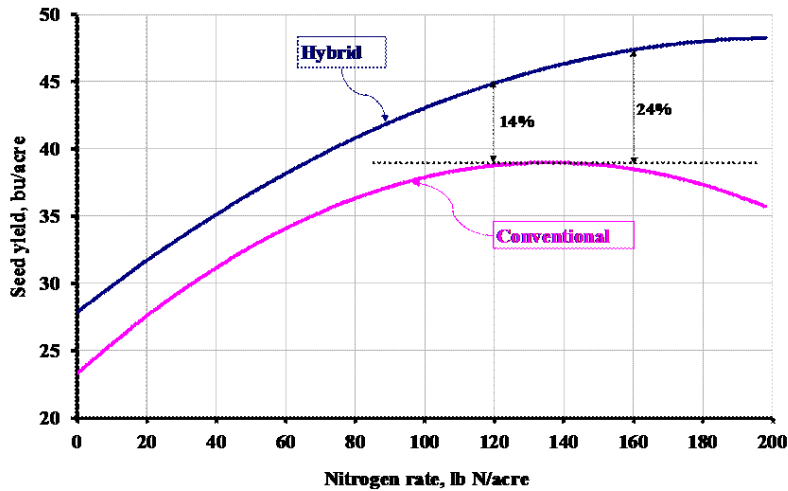
1. Failing to involve statistical considerations at the planning stage of the experiment.
2. Using improper experimental design or misusing a proper design.
3. Failing to use proper randomization procedures.
4. Using an improper size of an experiment.
5. Using improper experimental technique.
6. Using inappropriate error terms for testing or for calculating standard errors.
7. Failing to study patterns in data.
8. Depending excessively on one class of statistical analyses.
9. Misapplying multiple comparison procedures such as Duncan's new multiple range test.
10. Failing to report in the Materials and Methods section of the research report the experimental design and statistical procedures used.

I will illustrate how statistical procedures can be used and misused using two experiments of my own and one hypothetical one.

### Hybrid versus Conventional Canola

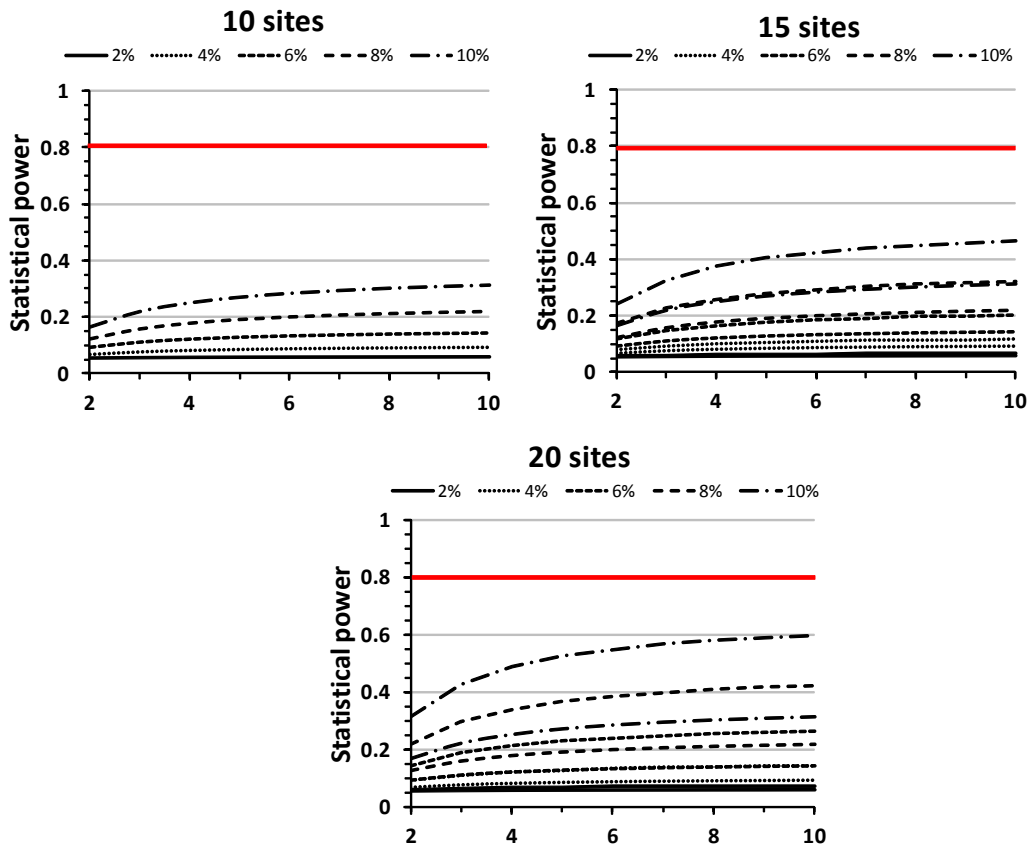
Two series of experiments were compiled into two publications (Karamanos et al. 2005; 2007). The experiment involved a large number (30) of regression experiments with twelve rates but no replication. The results converted to imperial units are summarized in Fig. 3. From the hybrid cultivar regression, a nitrogen (N) fertilizer rate of 180 kg N ha<sup>-1</sup> (160 lb/acre) that maximized yield and the corresponding SD (variance) was established. Then, a 'standard' mixed model ANOVA for the N fertilizer rate data (hybrid cultivar only and 180 kg N ha<sup>-1</sup>) was run to establish the proportions of variances ascribable to site, site x treatment, replicate within site, and residual random effects.

Then next phase was to run a mixed model simulation to test whether a 2–10% yield benefit from a product X can be detected for an optimally managed hybrid canola. The standard deviation derived from the first regression and the variance proportions from the second mixed model ANOVA were used to seed variance estimated for the mixed model simulation. For the simulations, the number of replication was varied, the sites were increased from 10 to 20, and treatment differences were increased from 2% greater than check to 10% greater than check.



**Figure 3.** Comparison of overall yield obtained for hybrid to that of conventional canola cultivars based on applied nitrogen rates.

A desired level statistical power of 0.8 (Peterman et al. 1990) was never achieved, even if a maximum amount of replicates and sites, and 10% benefit for product X were considered. In fact, a 4-replicate 20-site design was only able to achieve a statistical power of 0.47 (Fig. 4). Therefore, ability to detect the benefits at the upper end of curve becomes very difficult.



**Figure 4.** Statistical power for the hybrid canola experiments based on site numbers.

Statistical power analysis to achieve a power of 0.8 for linear slope by genetics interaction one would need hundreds of sites instead of 30-some when one considers data averaged across replicates (site by N rate by genetics {conventional and hybrid}). Hence, combining sites may not be desirable; rather, it might be better to assess power on a by-site basis because it becomes unmanageable when all sites are lumped together.

The difficulty in deriving benefits at the upper portion of the curves is illustrated using predictions and variance structure for the mixed model regression (random coefficient model) for these experiments (Fig. 5). The confidence limits show about 30% greater variability at the top of curves relative to lowest N rates.

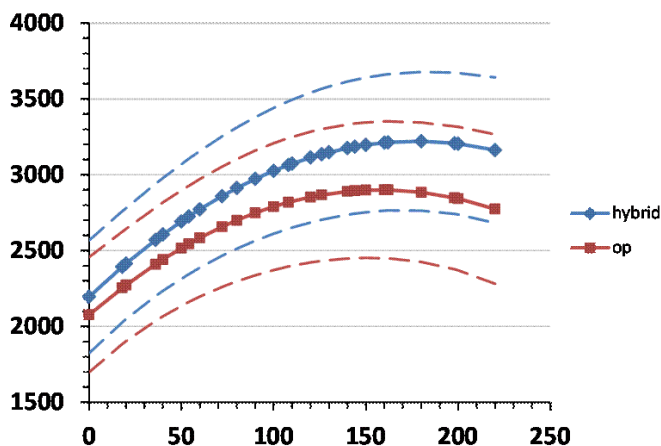


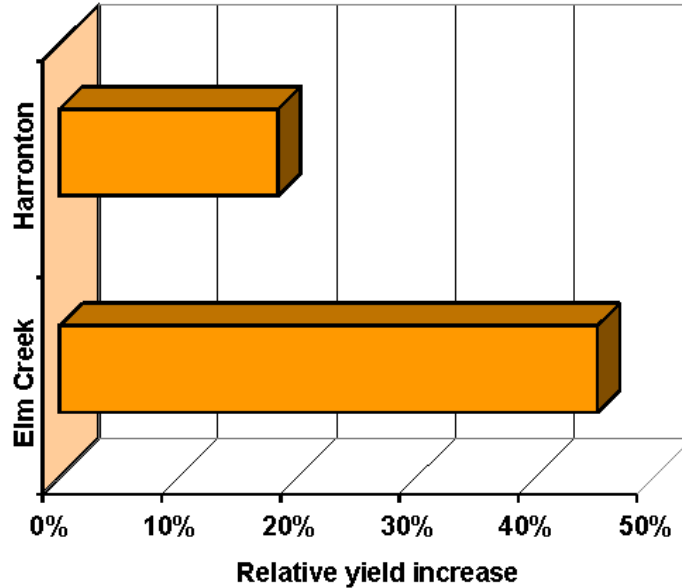
Figure 5. Confidence limits for hybrid and conventional (op) canola.

### The Two Penny (\$5.50 per acre) Experiment

Back in 2002, I presented two versions of the same data on a series of twenty experiments involving a control and a 2 ¢ treatment that, when converted on a per acre basis, it translates to \$5.50/acre (Karamanos and Flaten 2002a,b). The purpose of those presentations was to illustrate uses and misuses of statistics for the benefit of all those involved in agronomic research. One version included 9 of the 22 experiments that were seeded to canola (Table 2) without revealing the remaining 13 that were seeded to others crops. The second version included only two of the 9 canola experiments, where large responses were obtained, albeit only one statistically significant, without revealing either the rest of the 7 canola or 13 experiments with other crops (Fig. 6).

Table 2. New treatment helps canola beat the weather - MSSS

Location	Yield, bu/acre		Yield increase		Δ\$Y/1\$
	Control	Treated	bu/acre	%	
Red Deer	47.2	47.9	0.6	1.3	\$0.7
Wetaskiwin	50.3	50.6	0.2	0.4	\$0.2
Herronton	11	10.9	-0.1	-1.2	-\$0.2
Herronton	10.3	12.2	1.9	18.4	\$2.2
Balzac	33.8	35.5	1.7	5	\$2.1
Balzac	33.2	33.5	0.4	1.2	\$0.4
Choicelend	42	41.5	-0.5	-1.2	-\$0.2
Elm Creek	17.6	25.5	8	45.5	\$9.4
Miami	26.9	31.2	4.3	16	\$5.2
Average	30.3	32.1	1.9	9.6	\$2.25



**Figure 6.** treatment helps canola beat the weather - SC

In the first version (Table 2), the difference in the yield between the “two-penny” treatment and the control at one site (Elm Creek, Manitoba) was significant at 95% probability level ( $P < 0.05$ ). The difference in the remaining sites was not significant and overall the difference of 1.9 bu/acre of canola or 9.6% yield increase was below what our “experimental eye” (variance) could see. In version two (Fig. 6), a 18.4% yield increase, as spectacular as it appears was a result of the \$5.50 (2¢) treatment 1.9 bu/acre more than the control yield of 10.4 bu/acre, a drought situation where variability in the field increases randomly! However, only reporting a percentage yield increase can lead to a farmer believing that this can happen at any yield level. Independently of how well a one site one season experiment was conducted it could represent the one out of twenty chances to get the wrong answer!

Estimates of covariance parameters and means from a 'standard' mixed model are utilized to do a retrospective power analysis of the “two penny” experiment. From this analysis, the number of replicates required to achieve statistical power of 0.8 (Peterman et al. 1990) were calculated. With 20 replicates for canola one would have adequate statistical power to detect a difference that in actuality does not exist (Table 3)!

**Table 3.** Number of replicates required to identify statistically significant difference with acceptable statistical power ( $> 0.8$ ) with canola in the “two penny” experiment.

Crop	Check (bu/acre)	2 penny (bu/acre)	$\beta = 0.2$ (no. reps) <sup>a</sup>
barley	3.28	3.28	> 30
canola	1.72	1.82	20
wheat	2.28	2.29	> 30

<sup>a</sup>The number of replicates required to achieve statistical power of 0.8 ( $\beta = 0.2$ ) {Peterman (1990)}.

By accounting for residual heterogeneity among sites, canola means were adjusted (Littell et al. 2006) to provide correct or 'real' mean estimates. The corrected means are now identical and

not even close to statistically different (Table 4). Therefore, with the correct mixed model one can avoid type I errors (saying a difference is significant when it is not).

**Table 4.** Number of replicates required to identify statistically significant difference with acceptable statistical power (>0.8) with canola in the “two penny” experiment after accounting for heterogeneity among sites.

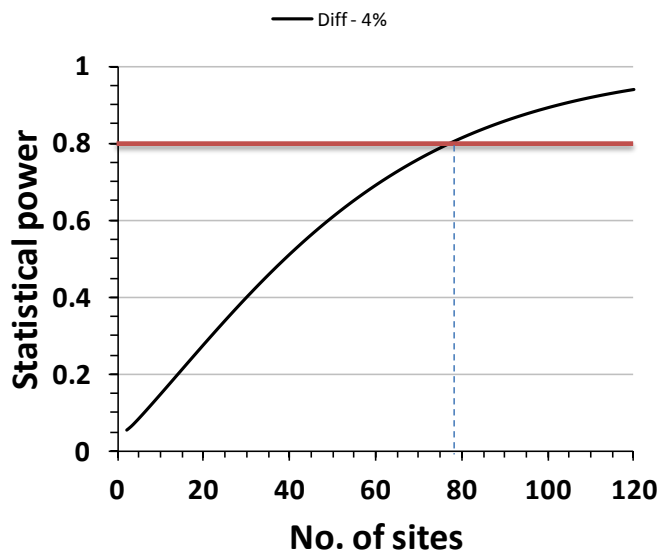
Crop	Check (bu/acre)	2 penny (bu/acre)	$\beta = 0.2$ (no. reps) <sup>a</sup>
barley	3.28	3.28	> 30
canola	1.77	1.77	> 30
wheat	2.28	2.29	> 30

<sup>a</sup>Number of replicates required to achieve statistical power of 0.8 ( $\beta = 0.2$ ) {Peterman (1990)}.

### The side-by-side comparison

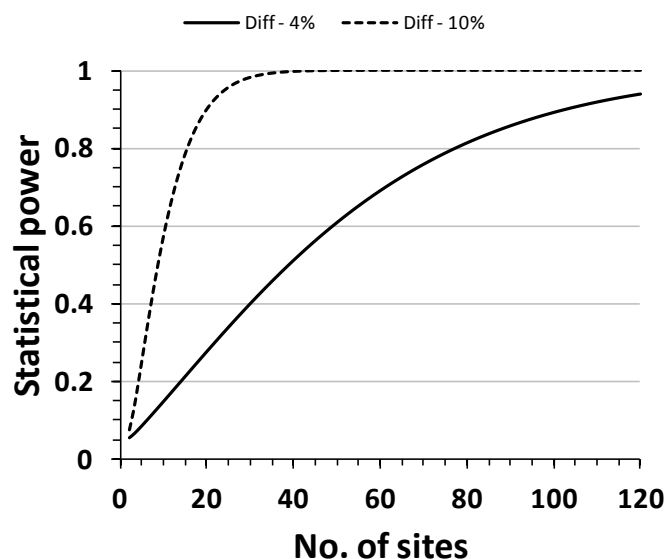
One hundred and forty five comparisons between a control and a treatment X were statistically analyzed using ANOVA (Table 5). The treatment P value was highly significant ( $P=0.0003$ ). The sites (location by year combinations) served as replicates in a randomized complete block design (RCBD) and the site by treatment variance was equal to the residual term. Strips at each site serve as experimental unit and treatments were randomly assigned to strips at each site. The statistical power, as expected was very strong (Fig. 7).

Treatment	Estimate t ha <sup>-1</sup>	SED	DDF*	$\Delta$ Yield t ha <sup>-1</sup>	bu/acre
Control	2.858	0.03254	137	0.120	1.79
Product	2.978				
Difference	4%				



**Figure 7.** Statistical power of a side-by-side one replication experiment.

A large enough number of sites (close to 80), which in this case are considered as replicates would provide the desired statistical power target of 0.8, which also signifies that with enough site (replicates) anyone can derive statistically significant differences. The question then becomes whether a relatively small percentage increase has practical significance or not, for if a 10% difference were to be considered to have more agronomic/economical importance, a power analysis with this difference using variance structure from the experiment shows that this design is exceedingly powerful (power = 0.9561, Fig. 8). Only 9 sites were needed to detect a 10% difference, and 15 sites were needed to achieve a desired statistical power of 0.8 for a 10% difference (Peterman et al. 1990).



**Figure 8.** Statistical power of the side-by-side one replication experiment when its variance structure is used to derive a 10% difference.

This example shows that a design with excessive precision can easily detect any difference that one wants.

### Acknowledgment

This presentation would not have been possible without the input and help by Dr. Craig Stevenson, who I deeply thank.

### References

- Bruulsema, T. 2009.** Recommendation development under 4R nutrient stewardship, in Proceedings North Central Extension-Industry Soil Fertility Conference. Volume 25. Des Moines, IA.
- Karamanos, R.E. 1992.** Commercial applications of root region management products - Will it happen? Presented at the Rhizosphere Management Seminar, March 19, Saskatoon, SK.
- Karamanos, R.E. and Flaten, D.2002a.** The "two penny" experiment. Proc. 45th Man. Soil Sci. Soc. Meeting, February 5-6, Man. Soil Sci. Soc., Winnipeg, MB.
- Karamanos, R.E. and Flaten, D. 2002b.** The \$5.50 per acre experiment. Proc. Soils and Crops 2002. February 21-22, Extension Division, Univ. of Saskatchewan, Saskatoon, SK.

- Karamanos, R.E., Goh, T.B. and Poisson, D.P. 2005.** Nitrogen, phosphorus and sulfur fertility of hybrid canola. *J. Plant Nutr.* 28: 1145-1161.
- Karamanos, R. E., Goh, T. B. and Flaten, D. N. 2007.** Nitrogen and sulphur fertilizer management for growing canola on sulphur sufficient soils. *Can. J. Plant Sci.* 87: 201–210.
- Littell, R.C, Milliken, G A., Stroup, WW, and Wolfinger, R D. 2006.** SAS System for Mixed Models (second edition) , Cary, NC: SAS Institute Inc.
- Nelson, L.A., and J.O. Rawlings. 1983.** Ten common misuses of statistics in agronomic research and reporting. *J. Agron. Educ.* 12:100–105.
- Peterman, R.M. 1990.** Statistical power analysis can improve fisheries research and management. *Can. J. Fish. Aquat. Sci.* 47: 2-16.
- Toft, CA. and Shea, P.J. 1983.** Detecting community-wide patterns: estimating power strengthens statistical inference. *Am. Nat.* 122: 618-625.