Ann. Zool. Fennici 46: 138–157 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 30 April 2009 © Finnish Zoological and Botanical Publishing Board 2009

Statistics: reasoning on uncertainty, and the insignificance of testing null

Esa Läärä

Department of Mathematical Sciences, P.O. Box 3000, FI-90014 University of Oulu, Finland (e-mail: esa.laara@oulu.fi)

Received 18 Feb. 2008, revised version received 20 Nov. 2008, accepted 29 Nov. 2008

Läärä, E. 2009: Statistics: reasoning on uncertainty, and the insignificance of testing null. — *Ann. Zool. Fennici* 46: 138–157.

The practice of statistical analysis and inference in ecology is critically reviewed. The dominant doctrine of null hypothesis significance testing (NHST) continues to be applied ritualistically and mindlessly. This dogma is based on superficial understanding of elementary notions of frequentist statistics in the 1930s, and is widely disseminated by influential textbooks targeted at biologists. It is characterized by silly null hypotheses and mechanical dichotomous division of results being "significant" ($P < 0.05$) or not. Simple examples are given to demonstrate how distant the prevalent NHST malpractice is from the current mainstream practice of professional statisticians. Masses of trivial and meaningless "results" are being reported, which are not providing adequate quantitative information of scientific interest. The NHST dogma also retards progress in the understanding of ecological systems and the effects of management programmes, which may at worst contribute to damaging decisions in conservation biology. In the beginning of this millennium, critical discussion and debate on the problems and shortcomings of NHST has intensified in ecological journals. Alternative approaches, like basic point and interval estimation of effect sizes, likelihood-based and information theoretic methods, and the Bayesian inferential paradigm, have started to receive attention. Much is still to be done in efforts to improve statistical thinking and reasoning of ecologists and in training them to utilize appropriately the expanded statistical toolbox. Ecologists should finally abandon the false doctrines and textbooks of their previous statistical gurus. Instead they should more carefully learn what leading statisticians write and say, collaborate with statisticians in teaching, research, and editorial work in journals.

Introduction

In ecology, as in many other scientific disciplines, statistical methods are extensively used in (a) *description* of variability in and summarization of empirical results obtained from experimental or observational studies, and in (b) *statistical inference* based on the observed data. In the first task the typical statistical tools are summary measures (like the mean, standard deviation, proportion, etc.), and tabular and graphical presentations of these or of the original data themselves. The second task involves assessment of random variation and uncertainty in the unknown parameters arising from the observed results. The common inferential tools are (i) testing of statistical hypotheses and computation of *P* values, (ii) interval estimation, and (iii) point estimation plus prediction.

Statistical inference belongs to the realm of inductive inference. It does not follow unequivocal rules like those of deductive logic as applied in pure mathematics. Just as various theories of inductive logic have been presented in philosophy of science, different schools of thought also exist in statistics. The two broad approaches to statistical inference may be labelled as (a) frequentist, and (b) Bayesian. In the first of them one can also distinguish the Fisherian and the Neyman-Pearson variants from each other.

Some popular textbooks on statistical methods targeted at biologists, like those of Sokal and Rohlf (1995), and Zar (1998), have obtained a highly influential and authoritative status among scientists in the biological disciplines all around the world. These texts tend to present a kind of hybrid of the Fisherian and Neyman-Pearson schools as the sole, monolithic and unquestioned doctrine of statistical inference. This dogma, prevalent in biological sciences, completely ignores the important differences between these two schools, the Bayesian paradigm (Gelman *et al.* 2003, McCarthy 2007), as well as the likelihood or evidential approach (Royall 1997). The presentation, even in the newest editions of these textbooks, gives a false impression as if no development has taken place in the principles and practices of statistics since the 1950s. Statistical analysis and inference are taught in them as a set of mechanical procedures governed by strict rules, and more or less blind obedience of nearly religious rituals is preached at the cost of independent thinking and common sense. These rituals concentrate around the notion of "statistical significance"; the magical event of a statistical test reaching a *P* value less than 5%, when typically testing a "null hypothesis" of exact zero difference. This is not what leading statisticians think and practice today, or did so already more than half a century ago. This ritual has long been condemned by many scientists in ecology and elsewhere, too, for reasons that will be discussed in this paper.

When criticizing the named textbooks it has to be said, though, that as handbooks they do contain a lot of useful and statistically sound technical material. However, the dominating message conveyed by them is the heavy emphasis given to testing "significance".

In this communication the prevalent statistical practice in ecology is critically reviewed. All of this has been written and said before in many other fora, so nothing really new is presented. Nevertheless, despite lively discussions of the pertinent issues in various ecological journals over the last two decades, the debate apparently has not reached all ecologists. Therefore, another comment may well be justified. Deeper philosophical issues concerning the nature of statistical analysis and inference and their role in scientific endeavour are discussed e.g. in Mayo (1996) and Taper and Lele (2004), the latter especially in the context of ecological research.

Examples of statistical malpractices

In this section, I shall illustrate some key features in the prevalent data-analytic paradigm as practiced in ecology as well as in other biological disciplines by examples, questions and answers, which pertain to very elementary and typical analyses. Although some aspects in these partly fictitious examples appear overly simplified and exaggerated, they do reflect real issues and problems in the current practice, and simplification serves to bring out the essential points.

Comparison of two groups

We first consider a simple two-group comparison. This is based on a real exam paper, targeted at students of biological sciences by a professor of ecology at University of D; it was found on the Internet after haphazard browsing.

The exam question was introduced as follows: "In an experiment I conducted on the effects of food availability to a small aquatic beetle (they eat fallen leaves, or leaf litter), I tested the hypothesis that the type of leaf litter in terms of the species of tree (red maple *vs.* white oak) from which the leaves came did not affect beetle growth rate. Raw data (μg day^{-1}) and the *t*-test I performed are shown below (Table 1)."

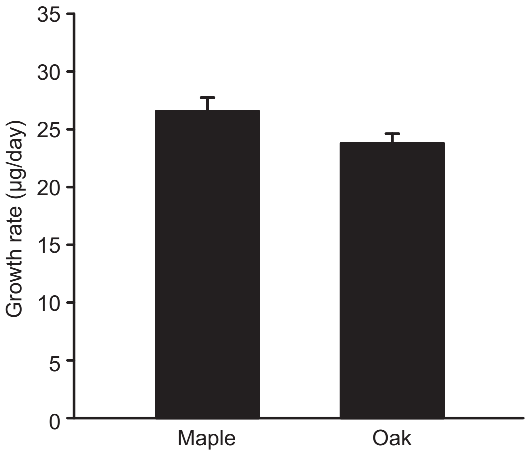

Fig. 1. Bar plot showing the mean and standard error of the mean (SEM) of growth rates by leaf type.

Graphics in two-group comparison

The error bar plot (Fig. 1) or "dynamite-plunger plot" (Freeman *et al.* 2008: pp. 9–10) showing the group means and their standard errors (SEM) is a commonly seen graphical illustration of a two-group comparison of quantitative variables in biological research articles. Yet, in textbooks of statistics written by statisticians, this graph appears to be rarely introduced, let alone recommended. In contrast it is criticized (e.g. Freeman *et al.* 2008) and is also viewed as "undesirable" by some medical journals in the *BMJ* group in their instructions to authors (e.g. http://jech.bmj.com/ifora/statadvice.pdf).

Table 1. Example data and output of statistical calculations as displayed by a spreadsheet calculation program.

Maple	25 30 22 27 33 26 29 24 23
Oak	24 22 21 26 22 25 22 29 23

t-test: two-sample assuming equal variances

	Leaves	
	Maple	Oak
Mean	26.555556	23.777778
Variance	12.777778	6.444444
Observations	9	9
Pooled variance	9.611111	
Hypothesized mean difference	0	
df	16	
t	1.900715	
$P(T \le t)$ one-tail	0.037755	
t critical one-tail	1.745884	
$P(\mathrm{T} \le t)$ two-tail	0.075510	
t critical two-tail	2.119905	

Question 1: What's wrong with this picture? Which kind of graphical presentation(s) would be better?

Answer 1: An error-bar plot with means and SEMs is overall a highly uninformative graphical presentation of results. Its *data/ink ratio* (Tufte 1983) is also very poor, meaning that a large amount of ink is spent to show only four numbers (two means and SEMs).

For *descriptive* purposes a *strip chart* or a *dot plot* of individual values is the best graph for small data sets, having an optimal data/ink ratio (Fig. 2). It shows the data, the whole data, and nothing but the data. The real variation within the groups will also be conveyed in contrast to the mean/SEM plot. Vertical comparisons are typically easier for human eyeballing than horizontal ones, favouring perhaps the right hand side version.

An alternative descriptive graph is a *box plot* (Fig. 3), which shows the median, quartiles, and extreme observations. This presentation is particularly good for "intermediate" group sizes. Again, the vertical version may be preferable.

In contrast to descriptive tasks, displaying only the group means and their SEMs in a graph is commonly believed to respond adequately to relevant *inferential* needs. The data/ink ratio of the graph could then be dramatically improved by substituting the voluminous pillars by simple dots placed at the levels of the mean values on the vertical axis. Even so, a substantial problem remains: how to read this type of a figure. The mean/SEM plot describes the 68% confidence intervals of single group means. Yet these items are not very useful when the quantities of interest are the *mean difference* and its precision, typically described by the 95% confidence interval (CI) of the difference. Some guidelines have been presented (e.g. Cumming & Finch 2005, Cumming *et al.* 2007) to help readers in interpreting graphs with error bars showing either

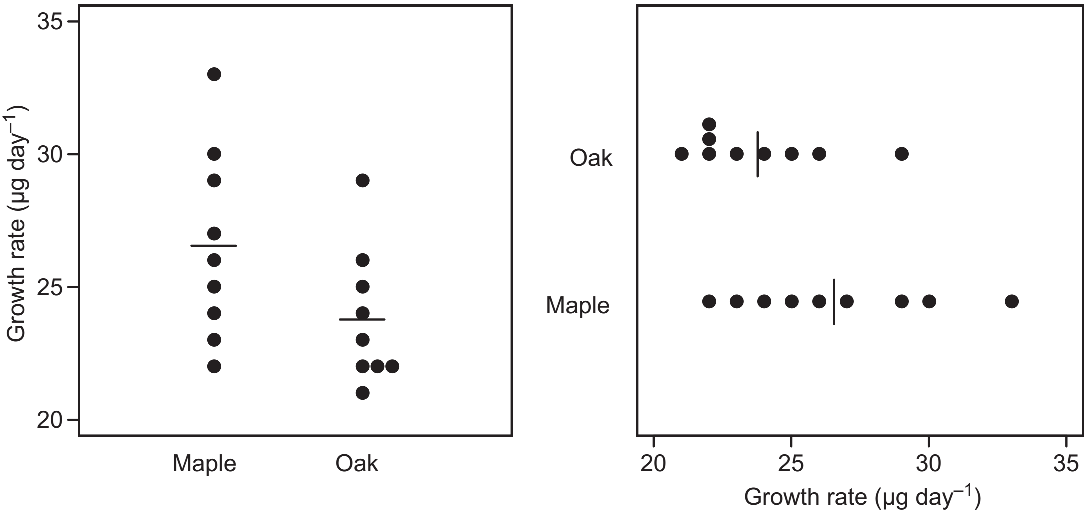

Fig. 2. Two versions of a strip chart showing growth rates by leaf type (group means indicated by short line segments).

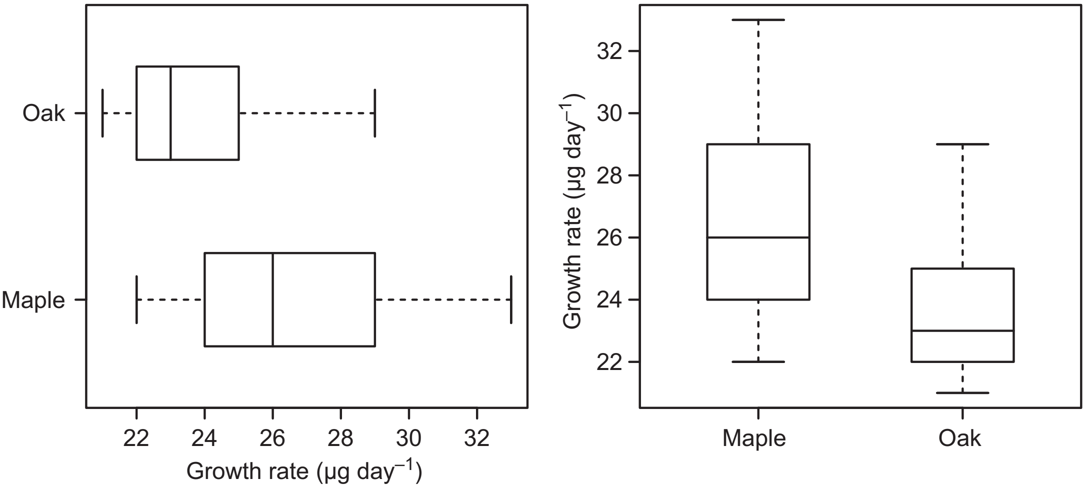

Fig. 3. Two versions of a box plot describing the distribution of growth rates by leaf type (median, quartiles, and range of values).

the SEMs or the 95% CIs about the single group means, when attempting to draw inferences about the mean difference. Unfortunately, due to the inherent limitations of these plots, these guidelines seem to be helpful at best in assessing whether the mean difference is "significant" or "non-significant" at the level 0.05. However, the precision in the estimated mean difference, concisely summarized in a confidence interval of this very contrast, is inadequately captured in graphs showing only group-specific statistics, but would need to be directly illustrated in a more suitable way (Cumming & Finch 2005). As suggested, e.g. by Freeman *et al*. (2008), it may often be most economical to limit reporting of the inferential statistics on the mean difference to the text or a table, and to use graphics only for descriptive purposes (like Fig. 2 or Fig 3). Nevertheless, in instances when a specific contrast is addressing the major question of scientific interest in a study, a more informative inferential graph than one giving a confidence interval at one confidence level (like 95%) only, would be a curve of the whole *confidence interval func-*

tion or *P-value function* (Cumming 2007), which shows the confidence interval simultaneously at all possible confidence levels.

Inference in two-group comparison

The actual exam question contained three items: "Based on all this information: (1) state my null and alternative hypotheses, (2) show the results of the *t*-test (the *t* value, degrees of freedom, and probability that the null hypothesis is correct), (3) draw the appropriate conclusion regarding food type and beetle growth."

Question 2: What's wrong in this exam question?

Answer 2: The starting point set out in item (1) will be commented on below in the context of the model solutions given. In item (2) the phrase "probability that the null hypothesis is correct" refers to a non-existent concept in frequentist statistics, in which the hypotheses have no such attribute as a probability of being correct. In Bayesian statistics, though, this kind of probability is a meaningful notion. It is apparent, however, that the author of the exam question is not a Bayesian.

Question 3: What items, relevant to statistical inference, are missing from the given output of statistical calculations?

Answer 3: The *point estimate* $\hat{\Delta}$ of the true mean difference Δ between the two types of leaves as well as the standard error of difference (SED) or confidence interval for it are missing. The values of these statistics are (μg day^{-1}): $\hat{\Delta} = 2.78$, SED = 1.46, 95% CI: [–0.3, +5.9]. The 95% CI shows a range of conceivable values for the true difference Δ with which the observed data (the group means and the within group variability) are in reasonable accordance. The width of CI reflects our uncertainty about Δ in light of the data. The importance of reporting point estimates and confidence intervals will be extensively discussed below.

The professor who gave this exam offered the following model solutions:

1. "Null: there is NO effect of leaf type on beetle growth. Alternative, or research hypothesis: there is an effect of leaf type on beetle growth (this is 2-tailed, as there is no specification as to which leaf type has a greater effect)."
2. "$t = 1.901$, df = 16, $P = 0.076$ (from 2-tailed test)."
3. "Since $P = 0.076$ is greater than our cut-off of 0.05, the null hypothesis is supported, and we must conclude that there is no effect of leaf type on beetle growth rates."

Question 4: What's wrong in this answer?

Answer 4: With regard to item (2), parts of the given answer would be OK, if only the question were rightly posed. However, all three items deserve to be commented:

1. This way of presenting the major question in statistical analysis is a representative case of the mindless *null ritual* (Gigerenzer 2004), widely practiced in biological research as well as in medicine and in behavioural sciences (Fidler *et al*. 2004a). The "silly null" (Anderson *et al*. 2001) stating an exactly zero difference, which in most instances is extremely unlikely *a priori*, and the broadest possible "alternative" stating "at least some non-zero difference" are most uninteresting and uninformative "hypotheses", and they provide no sensible starting point for decent research (Martínez-Abraín 2007).
2. The 2-tailed P value = 0.076 is not the "probability that the null hypothesis is correct". Instead, $P = 0.076$ stands for the probability of obtaining a contrast between the empirical group means that in absolute value would be at least as great as the observed mean difference in hypothetical replications of a similar experiment, *given that the null hypothesis stating* $\Delta = 0$ *were true*. The incorrect interpretation is one of the many common fallacies and illusions about P values (Gigerenzer 2004). In addition, these statistics pertaining to the silly null are extremely uninformative about the relevant *effect size*, i.e. the underlying quantitative contrast Δ in the mean responses between the two treat-

ments, and the imprecision or uncertainty associated with the point estimate of the true effect size based on the available data. In this regard the point estimate $\hat{\Delta}$ and its 95% confidence interval are much better. When they are reported, the t statistic and the P value convey no useful additional information on the effect size.

3. First, "our cut-off of 0.05" is completely arbitrary. Second, contrary to another persistent illusion — against which even trained statisticians are not necessarily immune (Lecoutre *et al.* 2003) — a "non-significant" result like this does not support H_0, even though it may be said to be *consistent with* H_0. Third, a conclusion "there is no effect of leaf type" is not a valid inference, because it is not logically implied by the observed results. Based on the confidence interval, the observed results actually provide equal, but quite weak, relative support for H_0: $\Delta = 0$ as they provide for Δ being 5.6 μg day^{-1}. If true, the latter value would imply over 20% higher growth rate with maple leafs. A good question then is, would a relative difference of this size be biologically important. If so, the empirical results are not excluding this possibility. — The best supported value for the unknown Δ, provided by the data, is its *maximum likelihood* (Cox 2006) point estimate $\hat{\Delta} = 2.8$ μg day^{-1}, the observed mean difference.

Sokal and Rohlf (1995), and Zar (1998) devote a disproportionately scanty attention to confidence intervals as compared with the emphasis given to "significance" testing. The computation of CIs for interesting treatment effects or group contrasts is covered, though. However, if any interpretation is given to the numerical intervals in the illustrative examples, the focus seems to be on whether the 95% CI covers the null value of Δ or not, i.e. whether the observed contrast is "significant" or not at the 5% level. Some other books targeted to biologists, like a recent text of Gotelli and Ellison (2004), tend to present CIs only for single group means but ignore them e.g. for the difference of means. In addition, Gotelli and Ellison (2004) introduces CIs for group means only in their chapter 3 on descriptive measures titled "Summary Statistics: Measures of Location and Spread" — a less proper context for explaining the meaning and use of inferential statistics.

Inference based on several studies — replication

Now we shall extend the example so that its results are complemented by those coming from other studies addressing the same research question.

Suppose that in addition to our professor at university D, research groups in four other universities A, B, C, and E conducted a similar experiment independently from each other (the group sizes being $n_1 = n_2 = 9$ in all). In all five studies a "non-significant" result was obtained (Table 2), and each research group made the same conclusion: "there is no effect of leaf type on beetle growth rates". Moreover, none of the groups managed to get their results published, because the referees and editors considered "non-significant" outcomes not worthy of publishing (Kotze *et al.* 2004). Nevertheless, a superficially logical consequence from the conclusions made in the separate studies would be: "The null hypothesis was confirmed, because all groups obtained a non-significant difference between maple and oak leaves".

Question 5. What's wrong with this reasoning?

Answer 5: "Non-significant" replications do not confirm H_0. In four of the five studies the observed mean difference appeared to favour maple leaf. The actual values for the point esti-

Table 2. Group means (μg day^{-1}), t statistics, and P values in five studies addressing the contrast between maple and oak leaves in the growth of beetles.

Study	Group means		*t*	*P*
	Maple	Oak		
A	24.1	22.4	1.17	0.26
B	25.1	21.7	1.82	0.09
C	24.8	24.9	–0.07	0.95
D	26.6	23.8	1.90	0.08
E	26.7	25.3	1.32	0.21

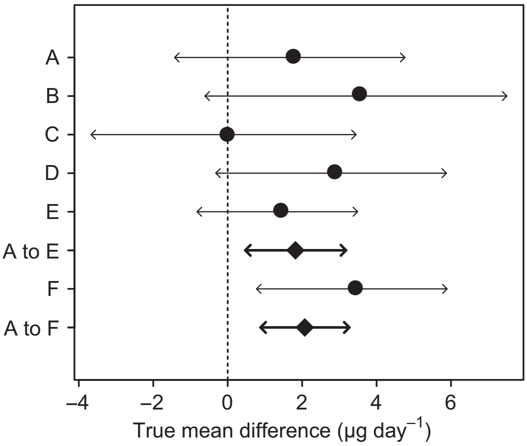

Fig. 4. Estimated mean differences and 95% confidence intervals from individual studies (bullets with thin arrow segments), and after pooling the first five and six studies, respectively (diamonds with thick line segments).

mates of the effect size, and its 95% confidence intervals were not reported in the table above. However, they may easily be recovered from the observed group means, t or P values and residual degrees of freedom using the well known statistical formulas plus computing procedures that provide exact quantiles of the t distribution. Facilities for computing densities, tail-area probabilities, and quantiles of several common distributions are nowadays easily available in numerous mathematical and statistical software packages. These make the traditional tables containing t, χ^2, and F quantiles for a limited number of "critical levels" outdated and unnecessary.

The point estimates and confidence limits of the mean differences from the individual studies can be graphically displayed in a *forest-plot* type of a diagram (Fig. 4) widely used in meta-analyses (Gurevitch & Hedges 2001, Freeman *et al.* 2008). Inspection of the graph suggests that the lower limits of these intervals were mostly closer to zero than the upper limits, all the latter being at least 3.4 μg day^{-1}. A simple meta-analysis of these five "non-significant" studies yields a pooled overall estimate for the mean difference: $\hat{\Delta} = 1.82$ with SE = 0.68, and 95% CI = [+0.47, +3.18] μg day^{-1}, respectively. Hence, these five studies together provide moderate evidence, which is actually favouring maple leafs. As argued by Kotze *et al.* (2004), publication of such apparently "negative" results would have been very desirable.

Independently of the above, a group at the Forest Research Institute (F) also conducted a similar experiment reporting the following results: "mean growth rates (μg day^{-1}) were 25.8 for maple, and 22.4 for oak; t = 2.78, P = 0.014*". Eventually, only this study was published, because it reached "statistical significance" — a classic instance of publication bias (Kotze *et al.* 2004)! A reviewer who was informed about the "non-significant" P values from the five unpublished studies commented on the study done in F: "This finding is in conflict with the unpublished results of other groups."

Question 6. What's wrong with this comment?

Answer 6: The study conducted at F with point estimate 3.4 μg day^{-1} is not at all conflicting but is well consistent with all the other studies. This should become clear by simple visual inspection of the point estimates and confidence intervals (Fig. 4). Pooled results of all six experiments, including that of F, were (in μg day^{-1}): overall mean difference = 2.07, SE = 0.60, and 95% CI = [+0.88, +3.27].

The evidence favouring maple leafs now becomes even stronger. The confidence interval for Δ based on the pooled analysis of all six studies gets ever narrower. Thus, the precision in the estimation is increased, and the uncertainty on the value of Δ is reduced.

Finally, a qualitative conclusion: "maple leafs are probably more effective than oak leaves" is justified without any statistics merely by leaning on the following classical principle: "Replication is a cornerstone of science. The question of interest is whether an effect size of a magnitude judged to be important has been consistently obtained across valid replications … [when] different investigators achieve similar results using different methods in different areas at different times. Whether any or all of the results are statistically significant is irrelevant. Replicated results make statistical significance testing unnecessary" (Johnson 1999).

Analysis of Variance (ANOVA)

An extension of a two-group comparison is that of several group means, for which the popular method is analysis of variance (ANOVA). Consider the following example, taken from a classic text on experimental design of Cochran and Cox (1957) and reanalyzed among others by Cox and Reid (2000) — an excellent modern introduction to the field. An agricultural field trial was conducted in order to address the effects of potash (5 levels) on the strength of cotton fibre in two blocks. The observations are plotted (Fig. 5), and the conventional ANOVA table is presented (Table 3).

The statistical model behind ANOVA is a linear regression model with group indicators serving as its explanatory terms (O'Hara 2009). Let Y_{ti} be the response in unit (plot) i under treatment t with μ_t ($t = 1, \ldots, T$) being the underlying "true" mean (theoretical expectation). The responses are assumed independent, normally distributed, and the error variance $\text{var}(Y_{ti}) = \sigma^2$ common for all t. The corresponding *global null hypothesis* H_0 is formulated

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_T$$

Typically the attention in the ANOVA table is devoted to the F statistic and the P value referring to the global null associated with the treatment factor.

In a popular text on experimental design Underwood (1997) characterizes the importance of the analysis of variance and F statistic as follows: "The analysis of variance is a procedure to allow a simple test of a logically complex null hypothesis. The hypothesis underlying all analyses of variance is that some difference is predicted to occur among the means. Having done the [ANOVA], the value of F obtained allows you to determine whether to reject or retain the null hypothesis of no difference among means."

If the above prescription were literally applied without regard to the continuous nature of the treatment factor, our naïve conclusion would be: "Since $F_{1,4} = 0.215$ was non-significant ($P > 0.05$), we conclude that potash has no effect on cotton strength."

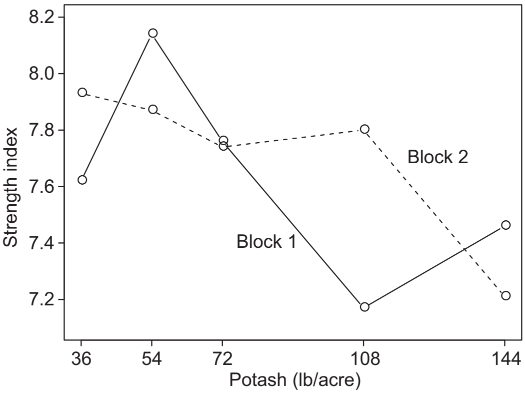

Fig. 5. Strength index of cotton by dose of potash in two blocks.

Question 7. Is this a valid inference, closing the story?

Answer 7: No, it's not.

Nearly all null hypotheses like this, assuming *exact equality* of means across several treatment levels, are most likely false on *a priori* grounds (Anderson *et al.* 2000), so the whole starting point of the F testing lacks meaning. The current view among mainstream statisticians on the global null, ANOVA table, and the F statistic is well expressed by Cox and Reid (2000, italics mine): "Such a [global] null hypothesis is very rarely of concern. [...] The SS for treatments is of importance primarily in connection with the computation of the residual SS, the basis for estimating the error variance. [...] *The main use of the [ANOVA] table is to provide an estimate of the standard error for assessing the precision of contrasts of the treatment means*. [...] With many statistical packages, the emphasis [...] is on the [ANOVA] table and the associated F-tests, which in nearly all cases are not the most useful information."

Table 3. ANOVA table from the experiment of potash on cotton strength.

Source	d.f.	SS	MS	*F*	*P*
Blocks	1	0.016	0.016	0.21	0.67
Treatments	4	0.552	0.138	1.85	0.28
Residual	4	0.298	0.075		

Hence, the only important quantities in this table are the residual mean square RMS = 0.075 and the associated residual degrees of freedom df = 4.

Because potash dose is a quantitative factor, interesting *contrasts* include at least the overall linear trend and the quadratic term representing curvature about this trend. A modern way of estimation of quantitative contrasts is based on fitting a corresponding regression model (O'Hara 2009) with linear, quadratic, and possibly higher order polynomial terms (orthogonalized) of potash. The standard errors and confidence intervals of the coefficients estimating these contrasts could be based on the RMS = 0.075 from the ANOVA model above, which represents the "pure error". However, if the variation about the linear trend can be assumed random, the error variance may be estimated from the residual MS of the regression model containing the linear term only. After this regression analysis, the estimated linear effect (slope) of potash on the stress index was –0.54 per 100 lb/acre (SE = 0.20, 95% CI [–0.98, –0.10]), which is "significant" ($P < 0.05$*)!

Type I error rate and multiple comparisons

What if the treatment factor were not quantitative but qualitative, having several unordered levels, and the F test for the global null were "significant"? What inferences on pairwise contrasts of any two different treatments are "admissible"? A substantial number of pages in many textbooks (e.g. Sokal & Rohlf 1995, Zar 1998) are spent on how to keep the "experiment-wise Type I error rate" at a desired level, and how to "correct" for the multiple comparison by methods like Bonferroni, Duncan, Dunn-Sidak, Newman-Keuls, etc. Yet, these concerns are completely ignored by e.g. Cox and Reid (2000).

Question 8. Should we bother about multiple comparisons?

Answer 8: To some extent yes, but not in any mechanistic way.

Below are quotations on how multiple comparison methods in the Anova context are viewed by some eminent statisticians: "Multiple comparison methods have no place at all in the interpretation of data" (Nelder 1971). "I have failed to find a single instance in which the Duncan test was helpful, and I doubt whether any of the alternative [multiple comparison] tests would please me better" (Finney 1988). "The ritualistic use of multiple-range tests — often when the null hypothesis is *a priori* untenable — is a disease" (Preece 1984).

In most instances of multiple comparisons it is entirely appropriate to report the point estimates and the nominal 95% confidence intervals for all interesting contrasts irrespective of whether specified *a priori* or considered afterwards. However, each confidence interval should be interpreted in light of relevant background knowledge pertaining to the particular contrast, and sound judgment is needed in these inferences to counter-balance the impact of formal mathematical argumentation.

To avoid misunderstanding, I wish to add that multiplicity does constitute a real problem in some exploratory studies addressing thousands of different associations at the same time, like when possible genetic factors determining certain phenotype values are screened by genome-wide scanning. However, even then the whole issue must be given more careful thought, rather than simply applying a blind prescription. *See* e.g. Ball (2007) for a more elaborate treatment of approaches and methods to respond to this highly topical challenge.

Normality tests and nonparametric methods

The mathematical theory pertaining to the F and t statistics says that the nominal properties of their sampling distribution (F and t distributions with given degrees of freedom, respectively) under H_0 are maintained, in case the observations can be regarded as random samples from underlying normal distributions with a common error variance.

A very common practice in ecological studies (and elsewhere too) is to perform a preliminary test on normality first. Based on the outcome of this test a decision will then be made: If a "significant" deviation from normality is found, a "non-parametric" or "distribution-free" analysis is chosen. With a "non-significant" outcome, one proceeds with the methods assuming a normal distribution.

Question 9. Is the practice of preliminary testing appropriate?

Answer 9: This practice is ill-founded, even paradoxical, because of the following reasons (Johnson 1995, Stewart-Oaten 1995): (1) Methods based on normality (like the *t*-test) are quite robust against typical deviations from normality, even in small samples. (2) With large group sizes the *central limit theorem* (CLT) implies approximate validity of normal-based methods. (3) Non-parametric methods are not so assumption-free as they often are advertised (Underwood 2009). They may actually be less robust against violations of assumptions (exactly similar shape and scale of the response distributions under different treatments) than the normal methods. (4) Modelling and estimation in more complicated settings becomes cumbersome with non-parametric methods. (5) The null hypothesis that the observations arise from a normal distribution is *a priori* almost certainly false anyway. (6) With large group sizes the normality tests are more sensitive to detect violations against the assumption on normality. Yet, it is precisely in these circumstances where CLT guarantees the approximate validity of the normal-based analysis. (7) With small group sizes the tests are not so powerful to reveal violations from the given assumptions. Therefore, a "non-significant" test result of normality does not justify a statement: "the data come from a normal distribution". Rather the result says: "the data do not provide sufficient evidence against the assumption of normality". Hence, with small groups the model assumptions may actually be more critical for the validity of the results than with large groups.

Null hypothesis significance testing (NHST)

The formulation of the exam questions, the proposed "model solutions", and the other examples are quite representative of the doctrine of *null hypothesis significance testing* (NHST), which is widely practiced in various fields of empirical sciences. This paradigm is characterized by heavy emphasis on zero null hypotheses (like "no effect of treatment", "zero difference between groups", and "null correlation between *X* and *Y*"), a tendency to dichotomize results into "significant" and "non-significant", mixing "statistically significant" with scientifically or practically important, and neglect of quantitative estimation and uncertainty assessment of interesting effects and contrasts.

Use of NHST in ecology

NHST has been applied in ecology since at least the 1950s. From a modest start the use of this procedure was rapidly expanded such that, for example, by 1970 about half of the original articles published in *Ecology* and *Journal of Ecology* reported results of significance tests (Fidler *et al*. 2004a). Recent surveys indicate that in the beginning of the new millennium the prevalence of NHST in ecologic journals has nearly reached a saturation level, such that clearly over 90% of the published papers contain them (*see* e.g. Anderson *et al*. 2000, Fidler *et al*. 2006, Stephens *et al*. 2007). However, Fidler *et al*. (2006) observed a slight decline in the use of significance tests in *Biological Conservation* and *Conservation Biology* from 2001–2002 to 2005 simultaneously with somewhat increased reporting of confidence intervals.

Key results from the survey of Anderson *et al*. (2000) on the occurrence of significance tests, *P* values, and inferences based on them in original articles published in *Ecology* during 1978–1997 and in *Journal of Wildlife Management (JWM)* during 1994–1998 may be summarized as follows: typically dozens of significance tests were reported (e.g. in *JWM* during 1996, the mean number of tests per article was 54, and ranged from 0 to 486). The greatest total numbers of tests in an annual volume were over 8000 (*Ecology* in 1991, *JWM* in 1996).

Faced with this heavy use of testing and massive amount of "significances" reported, a natural question arises: are all these tests really addressing biologically interesting and meaningful hypotheses? Is this practice a sign of sound theoretical thinking and mature science?

Anderson *et al.* (2000) observed that only few articles were found in which any of the tested H_0s could be considered plausible *a priori*. They commented: "A major research failing seems to be exploration of uninteresting or even trivial questions." Moreover, 47% of the *P* values and "significance" tests reported were "naked", i.e. they appeared alone without estimated means, differences, effects sizes, or associated measures of precision (like confidence intervals), or even the sign of the difference.

Why do significance tests maintain such a dominant position in the statistical practice as applied in ecology and elsewhere? The following points, given by Nester (1996), perhaps provide at least some superficial keys to understanding the present situation: "(a) [tests] appear to be objective and exact, (b) they are readily available and easily invoked in many commercial statistics packages, (c) everyone else seems to use them, (d) students, statisticians and scientists are taught to use them, and (e) journal editors and thesis supervisors demand them". In the following a brief historical excursion is given on how everything started and how the thinking and practices developed in statistics and ecology.

NHST — statistical technology of the 1930s

The doctrine of NHST is actually a mixture of two rival and even incompatible approaches to statistical testing: (a) significance testing developed by Fisher, and popularized in his two famous textbooks (Fisher 1925, 1935), and (b) hypothesis testing due to Neyman and Pearson, as presented in their seminal papers (Neyman & Pearson 1928, 1933). A couple of key features of this mixture are the following: (i) The idea of *null hypothesis* H_0 is from Fisher, but that of *hypothesis* (no null!) and *alternative hypothesis* from Neyman and Pearson. (ii) In Fisherian testing there is no alternative hypothesis. Hence there is no concept of *power* either, which on the other hand is essential in the Neyman-Pearson test theory. (iii) Fixed cut-off levels of "significance", like 0.05, 0.01 — advocated by Fisher — dominate interpretation and inference.

Nobody admits paternity of this mongrel. However, since World War II it has become very popular among mathematical statisticians first, and then it rapidly spread out to textbooks and teaching, being widely adopted in empirical sciences.

The advocacy of an arbitrary level of significance may be attributed to a blind spot of a genius. R. A. Fisher, the great statistician, geneticist and evolutionary biologist, wrote in his *Design of Experiments* (Fisher 1935): "Every experiment […] exist[s] only […] to give the facts a chance of disproving the null hypothesis. […] It is usual and convenient for experimenters […] to take 5 per cent as a standard level of significance, […] to ignore all results which fail to reach this standard."

This prescription had an enormous impact and unfortunate consequences. It led to *Worldwide Worship of Significance*: NHST was adopted and is still practiced like a pagan ritual by tens of thousands of researchers in many fields of science.

NHST viewed by present day statisticians

Since the 1930s, thinking and philosophy on the principles of statistics as well as its methodological tools have greatly developed. Other paradigms for statistical inference, including the likelihood or evidential approach (Royall 1997), and Bayesian statistics (Gelman *et al.* 2003, McCarthy 2007) have come on the scene. These alternatives have challenged the dominance of the frequentist school by pointing out some of its major deficiencies and paradoxes. Yet, outside statistics the primitive NHST dogma has survived almost intact for over half a century, and the subculture of statistical analysis and inference in many applied fields is more and more separated from mainstream statistics.

Since Berkson (1942) many eminent statisticians have explicitly criticised the NHST malpractice as applied in empirical sciences. The lack of emperor's clothes was explicitly revealed, e.g. by Frank Yates, a close friend and colleague of R. A. Fisher. On the 25th anniversary of Fisher's *Statistical Methods for Research Workers* Yates (1951) wrote: "the emphasis given to formal tests of significance […] has caused scientific research workers to pay undue attention to […] the results of the tests of significance […] too little to the estimates of the magnitude of the effects they are investigating […] the unfortunate consequence that [scientists] […] have often regarded the execution of a test of significance on an experiment as the ultimate objective […] the occasions […] in which quantitative data are collected solely with the object of proving or disproving a given hypothesis are relatively rare".

D. R. Cox, one of the most famous contemporary statisticians, has repeatedly commented on the NHST dogma (*see* also e.g. Cox 1958, 2006, Cox & Snell 1981), for example: "Overemphasis on tests of significance at the expense especially of interval estimation has long been condemned" (Cox 1977). "It is very bad practice to summarise an important investigation solely by a value of *P* […] The criterion for publication should be … not whether a significant effect has been found" (Cox 1982). "In one form or another this criticism has been repeated many times [since Yates 1951]" (Cox 2001).

These quotations reflect well the mainstream thinking on NHST among statisticians as well as their persistent frustration about its abuse (e.g. Nelder 1999). However, as many rival schools for statistical inference exist without unanimity concerning the principles, it is no wonder that there are also many statisticians, less famous though, who still recommend NHST as the primary tool for situations in which it is clearly inferior to other approaches and methods. Even worse, there is evidence that many trained statisticians also share certain common misconceptions and illusions about NHST (Lecoutre *et al*. 2003).

It is noteworthy that no reference to this discussion among professional statisticians that has endured over six decades is found in the texts of Sokal and Rohlf, or Zar, etc., even in their newest editions published in the 1990s.

NHST — criticized in ecological journals

Critical writings on the excessive and mindless use of NHST in ecological journals began to emerge about two decades ago (e.g. Jones & Matloff 1986, *see* also Anderson *et al*. 2000). In this regard ecology lagged behind e.g. psychological disciplines, in which the debate was initiated already in the 1960s (Rozeboom 1960), and medicine where it started somewhat later (e.g. Rothman 1978), as documented by Fidler *et al*. (2004a). This relative delay is understandable in light of the fact that testing "significance" was only introduced later into ecological research than to the behavioural and health sciences (Fidler *et al*. 2004a). During the 1990s awareness of the problems associated with NHST among ecologists increased (e.g. Yoccoz 1991, Stewart-Oaten 1995, Cherry 1998, Johnson 1999).

In the beginning of the 21st century the debate and discussion on issues of statistical inference, shortcomings of NHST, and alternative approaches has dramatically expanded and intensified in ecological journals (*see* e.g. Hobbs & Hilborn 2006, Nakagawa & Cuthill 2007, Stephens *et al*. 2007 for recent reviews). Among wildlife ecologists, the instrumental persons in these discussions have been D. H. Anderson, K. P. Burnham, and D. H. Johnson (e.g. Anderson *et al*. 2000, 2001, Johnson 1999, 2002). References to quotations by eminent statisticians and scientists on the issue are found from http://welcome.warnercnr.colostate.edu/~anderson/thompson1.html [compiled by W. Thompson].

Given the relative youth of this discussion among ecologists, it might be understandable that no reference to it could yet be found in the newest editions of e.g. Sokal and Rohlf (1995) and Zar (1998). These authors have not participated in this debate either.

The two major issues in these critics concern (i) the limitations of NHST as an inferential tool for answering ecologically relevant questions, and (ii) the poor understanding, persistent among users of NHST, about what the true logic and meaning of this procedure is.

The main inherent limitation of NHST is that at best it can only be used to evaluate the consistence or disagreement of observed data with a very simple statement, and nothing more. Situations in which these nulls are interesting are very rare: Most of the null hypotheses tested in ecological studies are silly nulls (Anderson *et al*. 2001), the exact truth of which is almost never plausible *a priori*. These nulls can always be "rejected" or "falsified" by a sufficiently large sample, so "testing" of such straw men does not really advance science (Anderson *et al*. 2000). In serious science, hypotheses, that possess any meaningful information content, are very different from banal propositions of claiming an exact zero effect or no difference, or from an empty "alternative" allowing non-zero effect of unspecified magnitude to whatever direction. Statistical null hypotheses are not such bold and informative conjectures in the Popperian sense which would merit severe efforts of refutation (Johnson 1999). Anderson *et al*. (2000) concluded that tests of statistical nulls have relatively little utility in science and are not a fundamental aspect of the scientific method.

Moreover, many statisticians have demonstrated (e.g. Berger & Sellke 1987, Royall 1997) how poor and misleading the *P* value actually is as a measure of evidence against the null hypothesis. The main problem is that the *P* value — tail-area probability — is based not only on the *observed* result (the data collected), but also on the more extreme and less likely, *unobserved* results (data sets that were never collected!). Hence, a *P* value is more of a statement about the events that never occurred than it is a concise statement of the evidence from the actually observed event (Anderson *et al*. 2000). More meaningful measures of relative evidence of any conceivable effect sizes (including zero) provided by empirical data are based on the *likelihood* (Royall 1997, Cox 2006) of these parameter values on the actually observed results; not on anything that was never observed.

Consequences of the NHST malpractice

Filling research reports with "significances", *P* values as such or with inequality signs, and/or stars may at best be viewed as a harmless sign of superstition among scientists. However, there are more serious aspects in the persistence of NHST: "Reporting of naked 'significances' or *P* values provides no information and is thus without meaning. Articles that employ silly nulls and statistical tests of hypotheses known to be false severely retard progress in our understanding of ecological systems and the effects of management programs. NHST doesn't give meaningful insights for conservation, planning, management, or further research" (Anderson *et al*. 2001).

As long as "non-significant" results are naively interpreted as providing evidence for "zero effect", "no change", "null correlation", etc., these may delay appropriate actions: "The consequences of accepting a false null hypothesis can be acute in conservation biology because endangered populations leave very little margin for recovery from incorrect management decisions" (Taylor & Gerrodette 1993). "For small populations, waiting for a statistically significant decline before instituting strong protection measures is often tantamount to a guarantee of extinction" (Fidler *et al*. 2006).

The credibility of scientists leaning on NHST in their argumentation is also at stake: "If ecologists are to be taken seriously by decision makers, they must provide information useful for deciding on a course of action, as opposed to addressing purely academic questions" (Johnson 1995).

Statistical reasoning and toolbox

In the context of statistical inference a major source of the existing confusion among scientists in various disciplines may well be caused by the language used: both that of mathematical formulas, containing Greek alphabets and other symbols, and the English words chosen, like "significant", "accept", "reject", "error", "power", "critical level", etc. In statistics, these words are just handy terms referring to certain abstract and purely technical concepts. Unfortunately, outside statistics they possess very strong connotations. Hence, scientists who are not used to statistical language, tend to take these notions literally and much more seriously than they ever deserve. The very word "inference" is particularly problematic because it may implicitly contain the notion of unequivocal rules and some predetermined algorithm always to be followed step by step from premises to conclusion, like in deductive logic. This holds not for statistical inference, nor for any other forms of inductive inference in science. Hence, a more appropriate term here would perhaps be *statistical reasoning* instead of statistical "inference".

Rules and judgments

Related to inference and reasoning, Stewart-Oaten (1995) presented the following remarks about rules and judgments in statistics: "Statistical analyses are based on a mixture of mathematical theorems and judgments based on subject matter knowledge, intuition, and the goals of the investigator. Textbooks and reviews, aiming for brevity and simplicity, blur the difference between mathematics and judgment. A folklore can develop, where judgments based on opinions become laws of what "should" be done. This can intimidate authors and readers, waste their time, and sometimes lead to analyses that obscure the information in the data rather than clarify it. Commonly obeyed rules are judgments with which it is reasonable to disagree."

In accordance with the above, even Sir Ronald Fisher himself wrote more thoughtfully, 20 years after his unfortunate statement quoted above: "[No] scientific worker has a fixed level of significance at which […] from year to year, and in all circumstances he rejects hypotheses; he rather gives his *mind* to each particular case in light of his *evidence* and his *ideas*" (Fisher 1956).

Also, even though Neyman and Pearson developed a very formal prescriptive theory for decision making based on strict test criteria and rules, they were actually much less dogmatic than many of their successors. Already in their seminal paper they wrote (Neyman & Pearson 1928, italics mine): "The process of *reasoning* […] is necessarily an individual matter, and we do not claim that the method which has been most helpful to ourselves will be of greatest assistance to others. It would seem to be a case where each individual must reason out for himself his own philosophy. […] The tests give no final verdict, but as tools help the worker […] to form his final decisions. What is of chief importance in order that a *sound judgment* may be formed is that *the method adopted, its scope and limitations, should be clearly understood*, […] we believe this often not to be the case […] it seems well to emphasize at the outset the *importance of careful thinking* in these matters."

These wise remarks apparently have not been reproduced in the various textbooks that maintain the common statistical folklore which has developed in ecology and other empirical sciences.

Some reasonable rules do exist, though, that can sincerely be offered for statistical analysis and reasoning. In a well-known text on ecological methodology Krebs (1998) presented 10 simple rules to be followed in empirical research. The first is about measurement, and I took the liberty to modify it to concern testing instead:

1. "Not everything that can be tested should be."

Rules 6 to 8 address specifically statistical issues, and they are:

6. "Never report an ecological estimate without some measure of its possible error."
7. "Be sceptical about the results of statistical tests of significance."
8. "NEVER confuse statistical significance with biological significance."

Statistical toolbox: alternatives to NHST

The key questions in statistical reasoning may be formulated: (1) what can be *learned* from the data obtained, or what *evidence* is obtained on the parameters of interest (like true underlying effect sizes) in quantitative terms, (2) what interpretations may be made based on the observations in the context of other relevant information, or how do the results modify our *prior* knowledge and beliefs.

Modern statistics does not contain only one single tool, like NHST ("Nail & Hammer — Sufficient for Timberwork"), for reasoning. It provides instead a rich toolbox from which, based on sound statistical insight and careful thinking, an appropriate tool may be chosen for any specific analysis problem at hand (Gigerenzer 2004). Significance tests have their place in the toolbox, because situations do exist where they remain useful (Cox 2006).

As has been repeatedly said, the very basic tools for statistical reasoning on the strength of associations and the sizes of differences and effects are provided by point estimates, their standard errors and associated confidence intervals (Johnson 1999, Anderson *et al*. 2001, Di Stefano & Fidler 2005, Cox 2006, Nakagawa and Cuthill 2007), especially in the context of well-thought statistical modelling (O'Hara 2009). In sufficiently complex models obtaining valid confidence intervals, or entire CI curves or confidence sets, is computationally much more challenging than getting a *P* value for the corresponding simple null hypothesis. Recent advances both in statistical theory and in software development have helped to make modern tools (e.g. profile likelihood, bootstrapping) more and more accessible to potential users.

Likelihood-based model fitting and information-theoretic measures in model selection have received considerable attention among ecologists, especially since the appearance of the books by Hilborn and Mangel (1997), and Burnham and Anderson (2002). These methods, when wisely applied, are highly useful in serious attempts to gain understanding on complex ecological processes using tools of statistical modelling.

Statistical modelling and inference based on the Bayesian paradigm has also recently been introduced to ecologists (*see* e.g. Ellison 1996, 2004, Clark 2005). A special issue of *Ecological Applications* (vol. 6, no. 4, 1996) was devoted to this theme, and a fresh textbook is published (McCarthy 2007). Bayesian inference incorporates relevant external information and knowledge to the interpretation of results obtained from an empirical study at hand. All uncertainty on parameters (like effect sizes) is expressed by probability distributions, in which probability is interpreted in terms of epistemic degrees of belief. The available external information is expressed by a *prior distribution* of the pertinent parameters. The information provided by the observed data from a single study is summarized by the *likelihood function* which is based on the statistical model postulated to the observable quantities. The information contained in the prior and the likelihood, respectively, is combined into the *posterior distribution* of the parameters.

Bayesian theory provides in principle a coherent probabilistic framework for inductive statistical inference. It is particularly powerful in the demanding task of synthesizing available relevant evidence contained in different kinds of research data coming from diverse sources. Progress in the computational tools implementing Markov chain Monte Carlo (MCMC) algorithms for approximating multidimensional posterior distributions (Gelman *et al*. 2003) have made complex Bayesian modelling computationally feasible.

The inherent subjectivity associated with formulating the prior distributions is the main reason for many statisticians and scientists to have strong reservations about the Bayesian paradigm. However, subjective elements are unavoidable in all statistical analysis and inference, whether Bayesian or not, starting from the often implicit model assumptions in all analyses. The virtue of the Bayesian approach is that it enables subjectivity to be explicitly and transparently incorporated to the analysis, and sensitivity of the final posterior inferences to various priors may be assessed.

Aside from these well-established mathematical constructs (models, likelihood, estimators, etc.), one should not forget the important tools of descriptive and explorative data analysis. Modern computing facilities and statistical software (like R, *see* R Development Core Team 2008) enable a rich repertory of graphical presentations, that are much more imaginative and informative than e.g. the ubiquitous dynamite-plunger plots. Graphical analysis will continue to have an essential role in statistical analysis and reasoning. To that end it is highly desirable that Excel and other commercial software designed for business graphics will soon disappear from the toolbox of scientists (Su 2008).

Different approaches and tools are complementary to each other. However, no real improvement in statistical analysis will take place, if the unthinking use of NHST is replaced by an equally mindless mechanical use of AIC or other information criteria upon model fitting by "user-friendly" statistical software. Confidence intervals can also be overused and misinterpreted — or not interpreted at all — as long as recommendations promoting their increased use are obeyed in an unthinking manner. In this regard Fidler *et al.* (2004b) have made an interesting observation associated with the recent change of statistical culture in medical journals. While the reporting of confidence intervals has increased in empirical research articles, the authors still do not seem to know how to interpret them.

Bayesian analysis in particular is extremely vulnerable to misuse, if adequate insight and skills are lacking from the user. In this regard it is very surprising to see in a recent book on ecological statistics by Gotelli and Ellison (2004), how Bayesian computations and inferential procedures are seemingly employed but on totally irrelevant quantities (like *F* ratio) and in such a meaningless way that has nothing to do with proper Bayesian statistics (*see* e.g. Gelman *et al.* 2003, McCarthy 2007). It is unfortunate that a text transmitting such fundamental misunderstandings has managed to be printed without a competent review being done beforehand, and is now doing a great disservice to the target audience. This is a pity, as certain parts in that book are sound, especially those covering aspects of study design.

The future of statistics in ecology

In spite of the increased public criticism of the mindless null ritual and of the growing awareness about useful alternatives that has taken place during the last decade, the change towards more appropriate practices seems to be slow. Fidler *et al.* (2006) found that, whereas the proportion of articles containing significance tests in *Conservation Biology* and *Biological Conservation* was 92% in 2000 to 2001, it was still as high as 78% in 2005. In contrast, the proportion of articles reporting confidence intervals in these journals remained modest, rising from 19% to only 26% in five years time. Moreover, significance testing was employed in 92% and 86% of the articles, respectively, in the volumes of 2005 of *Ecology* and *Journal of Ecology* (Fidler *et al.* 2006). In the survey of Stephens *et al.*, 90% to 100% of the papers in *Behavioural Ecology*, *Ecology Letters*, *Evolution*, and *Journal of Applied Ecology* still applied NHST in 2005.

Can something be done to stimulate faster progress in ecological statistics? A comparison with what has happened in the health sciences may be instructive in this regard.

Success story in health sciences

Fidler *et al.* (2004a) provide several reasons for the apparent success of a statistical reform in the medical journals, which has substantially diminished the overuse, abuse and misinterpretation of NHST, at the same time expanding the use of estimation and confidence intervals. Leaning very much on their extensive historical analysis, I will briefly comment the following three factors: (1) textbooks, (2) editorial policies, and (3) involvement of statisticians.

1. In contrast to Sokal and Rohlf, Zar, etc., the introductory (e.g. Campbell *et al.* 2007), and more advanced textbooks about statistics targeted at students and researchers in the health sciences, teach statistical inference less dogmatically and from a more balanced perspective. These texts pay special attention to the problems and limitations of NHST, and devote much more importance to the estimation of interesting contrasts and the assessment of their precision using confidence intervals, in accordance with common editorial policies (*see* below). Also, books on medical statistics are typically written by professional medical statisticians rather than physicians, apart from a few brilliant exceptions.
2. In the latter half of the 1980s some leading journals in medicine (like *BMJ*, *JAMA*, *Lancet*, *NEJM*) adopted more or less simultaneously an explicit editorial policy, according to which the use of mere significance testing in reporting statistical analyses was strongly discouraged and the use of confidence intervals promoted. Hundreds of journals have also joined to endorse the so called "Vancouver rules" i.e. the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals* (http://www.icmje.org), initiated in Vancouver in 1979. Since the 1988 edition these requirements have included a special section on statistics containing the following items

among others: "When possible, quantify findings and present them with appropriate indicators of measurement errors or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of *p* values, which fails to convey important quantitative information."

There are also other authoritative statements and guidelines for improving the quality and informativeness of medical research articles, like CONSORT for therapeutic and preventive trials, QUOROM for meta-analyses of trials, STROBE for observational studies, and MOOSE for their meta-analyses (*see* www.equator-network.org). All of these include explicit requirements for reporting confidence intervals for the interesting effect parameters as appropriate. Significance tests or *P* values are not specifically mentioned in these statements and guidelines; they are not expected to be used, nor completely banned either. Apart from these widely accepted recommendations on statistical reporting, many medical journals employ professional statisticians to review the statistical quality of the submitted manuscripts on a routine basis, in addition to subject matter reviewers.

All these determined editorial policies and guidelines have had an enormous positive impact on the quality of statistical analysis and presentation in journals of health sciences. The mindless NHST culture has not yet entirely vanished from medicine, but it is becoming more and more marginalized. However, there is still room for development in the statistical thinking of health scientists. It is desirable, for example, that "mandatory" reporting of confidence intervals would become increasingly accompanied by their more insightful interpretation than is currently the case (*see* e.g. Fidler *et al*. 2004b).

3. As already mentioned, professional medical statisticians have achieved an important role both as authors of textbooks on medical statistics and as reviewers of manuscripts submitted to medical journals. In addition statisticians are employed by medical schools as teachers for courses on statistics offered to medical undergraduates and postgraduates. Many of them are also plenipotentiary members of research teams, being especially indispensable in the design and analysis of clinical trials and large scale epidemiologic studies. Departments of biometry and biostatistics with permanent professors and other senior staff are not uncommon in medical schools.

Statistical reform in ecology — some recommendations

What then should be done in ecology and other biological disciplines in order to raise the level of statistical analysis and reasoning, which could simultaneously contribute to improved quality and usefulness of ecological research both from a purely scientific viewpoint but also from the pragmatic angle concerning the needs in environmental decision-making? The foregoing discussion, still largely based on the more detailed account of Fidler *et al*. (2004a), implies some suggestions:

1. Professional statisticians should be employed more than now by university departments of biology to have a substantial role in the teaching of statistics to biologists all the way from the elementary undergraduate level to most advanced postgraduate training.
2. Textbooks of statistical methods that are used in teaching and as methodological sources should preferably be written by statisticians with adequate experience and insight on biological research, at best co-authored by biologists possessing good knowledge of the approaches and methods of contemporary statistics. A good example of such a joint enterprise is the book written by Kenneth Burham, a statistician, and David Anderson, an ecologist (Burnham & Anderson 2002). Statisticians' involvement as members and co-authors in experimental and observational studies conducted by biological research teams should be increased, and their contribution should start already from the early design phase.
3. More work needs to be done in informing editors of ecological journals and collaborating with them in order to increase their awareness of appropriate statistical practices. Utilization of statisticians as reviewers

of submitted manuscripts should become a common editorial practice.

Statisticians and statisticians

When advocating a bigger role to statisticians in teaching, research, and editorial work, it is not my intention to claim that "if ecologists simply sought the advice of statisticians, all their statistical problems were solved". Statisticians are quite a heterogeneous species with a great variability at least with respect to their (i) training, (ii), understanding of fundamental inferential concepts, and (iii) involvement and experience with empirical applications.

At one extreme, in many universities, statistics is still being taught as a purely mathematical discipline having no touch with real-life research and data. Graduates from this kind of curriculum, who continue an academic career as mathematical statisticians in departments of mathematical sciences, most likely tend to teach statistics in the same spirit to new generations of students. Dogmatic attitudes concerning statistical testing and *P* values, and even usual misconceptions on them, are not uncommon among statisticians with such a background. Apart from brilliant exceptions, the contribution of pure mathematical statisticians can hence be quite thin and at worst even counter-productive considering the needs of ecologists or other researchers in biological sciences. Biologists possessing sound understanding about statistical principles and up-to-date knowledge of its methods may then be far more useful statistical experts for their colleagues.

At the other end there are universities where students of statistics are taught by teachers with a long track record of active personal involvement in applied research outside their own department. During their training the students achieve a fair theoretical knowledge basis and appropriate skills in statistical methodology, and at the same time they are exposed to realistic empirical applications in their courses and thesis work. Graduates coming from a curriculum like this form in principle a more promising pool for recruiting statisticians to the departments of biology than pure mathematical statisticians — let alone mathematicians or computer scientists without any formal training in statistics. However, it still takes many years of practical work experience for a newly graduated student to become an independent professional statistician possessing a deep enough insight on statistical reasoning plus adequate appreciation of the theory and practice of the subject matter field in which they are working. Such qualifications are fulfilled, for example, by those medical statisticians who have had an important role in the success story of the statistical reform in medicine and health sciences referred to above (Fidler *et al.* 2004a). The population of equally competent ecological statisticians is still quite small but will hopefully be growing in the near future.

Conclusion

The popularity of the dogma of null hypothesis significance testing is one of the mysteries and curiosities of 20th century science. It provides an ample opportunity with interesting topics for serious research in the history and sociology of science. We may nowadays laugh to the long-lasting practicing of alchemy over several centuries. Yet, how different are the primitive illusions and fantasies associated with NHST actually from the belief of the philosopher's stone?

From an applied statisticians' perspective the ritualistic practice of "analysing" empirical research data leaning on the outdated paradigm of NHST, as represented by Sokal and Rohlf, Zar and other similar texts, is a mockery of statistics and statistical science (Nelder 1999) of modern days. It is a subculture so much separated from the true statistics, biostatistics, or biometry of today, that such data-analytic activity actually deserves to be called by a completely different name, for example "significatistics".

The following appeal from a group of students of ecology offers, however, some grounds for optimism concerning the future of statistics in ecology: "We, as students […] encourage academic institutions […] to introduce students to various statistical approaches so we can make well-informed decisions on the appropriate use of statistical tools in wildlife and ecological research projects. […] We do not ask for the development of cookbook of statistical methods. […] We are not asking that the academic advisors be statistical gurus, but […] encourage [them] to become familiar with the statistical approaches available […] and thus decrease bias towards one approach. Null hypothesis significance testing is likely to persist as the most common statistical analysis tool […] until academic institutions and student advisors change their approach and emphasize a wider range of statistical methods" (Butcher *et al*. 2007).

Hopefully biologists will soon abandon their previous gurus, start listening instead what is being taught by leading representatives of contemporary mainstream statistics (e.g. Cox & Snell 1981, Cox 2006), and attempt to increase genuine collaboration with competent professional statisticians with interest and experience in empirical research. Statistics of the 21st century has a lot to offer for ecologists.

Acknowledgements

The author wants to express sincere thanks to Dr. Fiona Fidler, Dr. Bob O'Hara and an anonymous reviewer, whose thorough reading of the original manuscript and valuable comments substantially helped to improve the presentation. This paper was written while the author enjoyed the Research Grant for Senior Scientist No. 120146 from the Academy of Finland.

References

Anderson, D. R., Burnham, K. P. & Thompson, W. L. 2000: Null hypothesis testing: Problems, prevalence, and an alternative. — *Journal of Wildlife Management* 64: 912–923.

Anderson, D. R., Link, W. A., Johnson, D. H. & Burnham, K. P. 2001: Suggestions for presenting the results of data-analysis. — *Journal of Wildlife Management* 65: 373–378.

Ball, R. D. 2007: Statistical analysis and experimental design. — In: Oraguzie, N. C., Rikkerink, E. H. A., Gardiner, S. E. & De Silva, H. N. (eds.), *Association mapping in plants*: 133–209. Springer, New York.

Berger, J. O. & Sellke, T. 1987: Testing a point null hypothesis: the irreconcilability of *P* values and evidence. — *Journal of the American Statistical Association* 82: 112–122.

Berkson, J. 1942: Tests of significance considered as evidence. — *Journal of the American Statistical Association* 37: 325–335.

Burnham, K. P. & Anderson, D. R. 2002: *Model selection and multi-model inference*, 2nd ed. — Springer-Verlag, New York.

Butcher, J. A., Groce, J. E., Lituma, C. M., Cocimano, M. C., Sánchez-Johnson, Y., Campomizzi, A. J., Pope, T. L., Reyna, K. S. & Knipps, A. C. S. 2007: Persistent controversy in statistical approaches in wildlife science: a perspective of students. — *Journal of Wildlife Management* 71: 2142–2144.

Campbell, M. J., Machin, D. & Walters, S. J. 2007: *Medical statistics: a textbook for the health sciences*, 4th ed. — John Wiley and Sons, Chichester.

Cherry, S. 1998: Statistical tests in publications of the Wildlife Society. — *Wildlife Society Bulletin* 26: 947–953.

Clark, J. S. 2005: Why environmental scientists are becoming Bayesians. — *Ecology Letters* 8: 2–14.

Cochran, W. G. & Cox, G. M. 1957: *Experimental design*. — John Wiley and Sons, New York.

Cox, D. R. 1958: Some problems connected with statistical inference. — *Annals of Mathematical Statistics* 29: 357–372.

Cox, D. R. 1977: The role of significance tests. — *Scandinavian Journal of Statistics* 4: 49–70.

Cox, D. R. 1982: Statistical significance tests. — *British Journal of Clinical Pharmacology* 14: 325–331.

Cox, D. R. 2001: Another comment on the role of statistical methods. — *British Medical Journal* 322: 231.

Cox, D. R. 2006: *Principles of statistical inference*. — Cambridge University Press, Cambridge.

Cox, D. R. & Reid, N. 2000: *The theory of the design of experiments*. — Chapman & Hall/CRC, Boca Raton.

Cox, D. R. & Snell, E. J. 1981: *Applied statistics: principles and examples*. — Chapman and Hall, London.

Cumming, G. 2007: Inference by eye: pictures of confidence intervals and thinking about levels of confidence. — *Teaching Statistics* 29: 89–93.

Cumming, G. & Finch, S. 2005: Inference by eye: confidence intervals, and how to read pictures of data. — *American Psychologist* 60: 170–180.

Cumming, G., Fidler, F. & Vaux, D. L. 2007: Error bars in experimental biology. — *The Journal of Cell Biology* 177: 7–11.

Di Stefano, J. & Fidler, F. 2005: Effect size estimates and confidence intervals: an alternative focus for the presentation and interpretation of ecological data. — In: Burk, A. R. (ed.), *New trends in ecology research*: 71–102. Nova Science, New York.

Ellison, A. 1996: An introduction to Bayesian inference for ecological research and environmental decision-making. — *Ecological Applications* 6: 1036–1046.

Ellison A. 2004: Bayesian inference in ecology. — *Ecology letters* 7: 509–520.

Fidler, F., Cumming, G., Burgman, M. & Thomason, N. 2004a: Statistical reform in medicine, psychology and ecology. — *The Journal of Socio-Economics* 33: 615–630.

Fidler, F., Burgman, M., Cumming, G., Buttrose, R. & Thomason, N. 2006: Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. — *Conservation Biology* 20: 1539–1544.

Fidler, F., Thomason, N., Cumming, G., Finch, S. & Leeman, J. 2004b: Editors can lead researchers to confidence intervals but can't make them think: statistical reform lessons from medicine. — *Psychological Science* 15: 119–126.

Finney, D. J. 1988: Was this in your statistics textbook? III. Design and analysis. — *Experimental Agriculture* 24: 421–432.

Fisher, R. A. 1925: *Statistical methods for research workers*. — Oliver and Boyd, Edinburgh.

Fisher, R. A. 1935: *Design of experiments*. — Oliver and Boyd, Edinburgh.

Fisher, R. A. 1956: *Statistical methods and scientific inference*. — Oliver and Boyd, Edinburgh.

Freeman, J. V., Walters, S. J. & Campbell, M. J. 2008: *How to display data*. — Blackwell Publishing, Oxford.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. B. 2003: *Bayesian data analysis*, 2nd ed. — *Chapman & Hall/CRC, Boca Raton*.

Gigerenzer, G. 2004: Mindless statistics. — *The Journal of Socio-Economics* 33: 587–606.

Gotelli, N. J. & Ellison, A. M. 2004: *A Primer of ecological statistics*. — Sinauer, Sunderland MA.

Gurevitch, J. & Hedges, L. V. 2001: Meta-analysis: combining the results from independent experiments. — In Scheiner, S. M. & Gurevitch, J. (ed.), *Design and analysis of ecological experiments*, 2nd ed.: 347–369. Oxford University Press, Oxford.

Hilborn, R. & Mangel, M. 1997: *The ecological detective. Confronting models with data*. — Princeton University Press, Princeton.

Hobbs, N. T. & Hilborn, R. 2006: Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. — *Ecological Applications* 16: 5–19.

Johnson, D. H. 1995: Statistical sirens: The allure of nonparametrics. — *Ecology* 76: 1998–2000.

Johnson, D. H. 1999: The insignificance of statistical significance testing. — *Journal of Wildlife Management* 63: 763–772.

Johnson, D. H. 2002: The role of hypothesis testing in wildlife science. — *Journal of Wildlife Management* 66: 272–276.

Jones, D. & Matloff, N. 1986: Statistical hypothesis testing in biology: a contradiction in terms. — *Journal of Ecologic Entomology* 79: 1156–1160.

Kotze, D. J., Johnson, C. A., O'Hara, R. B., Vepsäläinen, K. & Fowler, M. S. 2004: Editorial: The Journal of Negative Results in Ecology and Evolutionary Biology. — *Journal of Negative Results* 1: 1–5.

Krebs, C. J. 1998: *Ecological methodology*, 2nd ed. — Addison-Wesley, Menlo Park, CA.

Lecoutre, M.-P., Poitevineau, J. & Lecoutre, B. 2003: Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. — *International Journal of Psychology* 38: 37–45.

Martínez-Abraín, A. 2007: Are there any differences? A non-sensical question in ecology. — *Acta Oecologica* 32: 203–206.

Mayo, D. 1996: *Error and the growth of experimental knowledge*. — The University of Chicago Press, Chicago.

McCarthy, M. A. 2007: *Bayesian methods for ecology*. — Cambridge University Press, Cambridge.

Nakagawa, S. & Cuthill, I. C. 2007: Effect size, confidence interval and statistical significance: a practical guide for biologists. — *Biological Reviews* 82: 591–605.

Nelder, J. A. 1971: Discussion. — *Journal of The Royal Statistical Society B* 33: 244–246.

Nelder, J. A. 1999: From statistics to statistical science. — *The Statistician* 48: 257–269.

Nester, M. R. 1996: Applied statistician's creed. — *Applied Statistician* 45: 401–410.

Neyman, J. & Pearson, E. S. 1928: On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. — *Biometrika* 20: 175–240.

Neyman, J. & Pearson, E. S. 1933: On the problem of the most efficient test of statistical hypotheses. — *Philosophical Transactions of the Royal Society of London, Series A* 231: 289–337.

O'Hara, R. B. 2009: How to make models add up — a primer on GLMMs? — *Annales Zoologici Fennici* 46: 124–137.

Preece, D. A. 1984: Biometry in the Third World: science, not ritual. — *Biometrics* 40: 519–523.

R Development Core Team 2008: R: a language and environment for statistical computing. — R Foundation for Statistical Computing, Vienna, Austria, available at http://www.R-project.org.

Rothman, K. J. 1978: A show of confidence. — *New England Journal of Medicine* 299: 1362–1363.

Royall, R. M. 1997: *Statistical evidence: the likelihood paradigm*. — Chapman & Hall, London.

Rozeboom, W. W. 1960: The fallacy of the null hypothesis significance test. — *Psychological Bulletin* 57: 416–428.

Sokal, R. R. & Rohlf, N. J. 1995: *Biometry: the principles and practice of statistics in biological research*, 3rd ed. — W.H. Freeman and Company, New York.

Stephens, P. A., Buskirk, S. W. & Martínez del Rio, C. 2007: Inference in ecology and evolution. — *Trends in Ecology and Evolution* 22: 192–197.

Stewart-Oaten, A. 1995: Rules and judgments in statistics: three examples. — *Ecology* 76: 2001–2009.

Su, Y.-S. 2008: It's easy to produce chartjunk using Microsoft® Excel 2007 but hard to make good graphs. — *Computational Statistics and Data Analysis* 52: 4594–4601.

Taper, M. L. & Lele, S. R. (eds.) 2004: *The nature of scientific evidence: statistical, philosophical, and empirical considerations*. — The University of Chicago Press, Chicago.

Taylor, B. & Gerrodette, T. 1993: The uses of statistical power in conservation biology: the vaquita and northern spotted owl. — *Conservation Biology* 7: 489–500.

Tufte, E. 1983: *The visual display of quantitative information*. — Graphics Press, Cheshire.

Underwood, A. J. 1997: *Ecological experiments: their logical design and interpretation using analysis of variance*. — Cambridge University Press, Cambridge.

Underwood, A. J. 2009: Components of design in ecological field experiments. — *Annales Zoologici Fennici* 46: 93–111.

Yates, F. 1951: The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. — *Journal of the American Statistical Association* 46: 19–34.

Yoccoz, N. G. 1991: Use, overuse, and misuse of significance test in evolutionary biology and ecology. — *Bulletin of the Ecological Society of America* 72: 106–111.

Zar, J. F. 1998: *Biostatistical analysis*, 4th ed. — Prentice-Hall, Upper Saddle River, NJ.

This article is also available in pdf format at http://www.annzool.net