Ardeola 52(2), 2005, 377-387

FORUM

Forum es una sección que pretende servir para la publicación de trabajos de temática, contenido o formato diferentes a los de los artículos y notas breves que se publican en *Ardeola*. Su principal objetivo es facilitar la discusión y la crítica constructiva sobre trabajos o temas de investigación publicados en Ardeola u otras revistas, así como estimular la presentación de ideas nuevas y revisiones sobre temas ornitológicos actuales.

*The **Forum** section of* Ardeola *publishes papers whose main topic, contents and/or format differ from the normal articles and short notes published by the journal. Its main aim is to serve as a lighter channel for discussion and constructive criticism on papers or reseach lines published either in* Ardeola *or elsewhere, as well as to stimulate the publication of new ideas and short revisions on current ornithological topics.*

CAN ORNITHOLOGY ADVANCE AS A SCIENCE RELYING ON SIGNIFICANCE TESTING? A LITERATURE REVIEW IN SEARCH OF A CONSENSUS

Alejandro MARTÍNEZ-ABRAÍN*[1] & Daniel ORO*

SUMMARY.—*Can ornithology advance as a science relying on significance testing? A review in search of a consensus.*

There is a sense of frustration among ornithologists, especially among those dealing with conservation biology, regarding the usefulness of research results for decision-making. Similar problems affect the social and behavioural human sciences and their failures have been partially linked with the strong dependence of data analysis on the statistical testing of null hypothesis. We review the literature on reasoning against and in favour of null hypothesis testing, in search of some consensus procedure for proper data analysis, so that knowledge acquisition is feasible and decision-making policies are useful. We conclude that a consensus on appropriate data analysis could be reached if we focus on precision of the estimates and biological relevance instead of on significance. This implies an important change in our way of thinking, promoting confidence intervals for the difference of effects rather than p-values. The alternative approach of Bayesian statistics can also be of great help in the decision-making process typical of applied ornithology as it provides a measure of the evidence of the effects and can solve complex models with straightforward procedures. We suggest that the slow progress of theory accumulation in ornithology and the low reliability of results for decision making may be due, to a large extent, to the way we analyze data rather than to the nature of the topics approached by ornithologists.

Key words: alternatives, Bayesian statistics, confidence intervals, decision-making, knowledge accumulation, null hypothesis, power analysis, P-values, significance testing.

* IMEDEA (CSIC-UIB), C/Miquel Marqués 21, E-07190 Esporles, Mallorca, Spain.
[1] Corresponding author: a.abrain@uib.es

RESUMEN.—*¿Puede la ornitología avanzar como una cienca descansando sobre la significación de los análisis? Una revision en busca de consenso.*

Existe un sentimiento de frustuación entre los ornitólogos, especialmente entre aquellos dedicados a la biología de la conservación, sobre la falta de utilidad de los resultados de las investigaciones científicas para los gestores del medio ambiente. Problemas semejantes afectan a las ciencias sociales y del comportamiento humano, y sus fracasos han sido parcialmente relacionados con la dependencia fuerte que tienen los análisis de los datos de la comprobación estadística de la hipótesis nula. En este trabajo revisamos la literatura existente y exponemos argumentos en favor y en contra de la comprobación de la hipótesis nula. Pretendemos mostrar métodos consensuados para analizar adecuadamente los datos, que permitan el progreso del conocimiento y la posibilidad de tomar decisiones políticas por parte de los gestores del medio ambiente. Se concluye que se puede alcanzar un consenso en como analizar los datos de forma apropiada, si nos centramos en la precisión de las estimas y en su relevancia biológica, en lugar de centrarse en su significación estadística. Esto conlleva un importante cambio en nuestra forma de pensar, fomentando los intervalos de confianza de las diferencias de los efectos en lugar de los valores de significación (valores de la P). La aproximación alternativa de los estadísticos "Bayesianos" puede ser también de gran ayuda en los procesos de toma de decisiones típicas de la ornitología aplicada, ya que nos ofrece una medida de la existencia de los efectos y puede resolver modelos complejos por medio de procedimientos directos. Sugerimos que el lento progreso teórico de la ornitología y la escasa fiabilidad de los resultados para los gestores ambientales pueden ser debidos, en gran parte, a la forma en que nosotros analizamos los datos más que en la naturaleza de los temas de investigación que son afrontados por los ornitólogos.

Palabras clave: alternativas, estadistica "Bayesiana", intervalos de confianza, gestores ambientales, acumulación de conocimiento, hipótesis nula, analisis de poder, valores de P, comprobación significativa.

INTRODUCTION

Debates on the misinterpretation and misapplication of statistics in biology are useful and necessary. We all know that the assumptions of the tests (e.g. sample independence, homoscedasticity, normal distribution of errors, etc) are commonly ignored, that the cut-off value of α is a convention and that significance testing is often misused and misinterpreted (Salsburg, 1985; Yoccoz, 1991; Mulaik *et al.*, 1997; see James & McCullogh, 1990 for a complete analysis of misapplications of multivariate analysis).

However, all debates dealing with specific statistical topics are minor compared to the one dealing directly with the utility of statistical significance testing. Null hypothesis statistical testing (NHT hereafter) is the pervasive paradigm behind knowledge acquisition in most sciences, including ornithology. It was first popularized by the English statistician R. A. Fisher (Fisher, 1951) and further developed through the Neyman-Pearson approach that proposed establishing a significance level (i.e. a risk level) in advance of the data collection, and introduced the concept of alternative hypotheses (Quinn & Keough, 2004).

Nowadays, most statistical tests in ornithology, from simple comparisons between means to complex linear and non-linear models, are based on the NHT rationale. However, a debate on its utility started within a few decades of its popularization, was revisited during the 70s, and was recently reopened during the 90s mainly by Cohen (1994) and Schmidt (1996). This discussion has taken place mainly among social and behavioural sciences, in specialized journals or books of psychology, education and sociology (see *e.g.* Oakes, 1986). Psychologists approached this topic to try to recover from the generalized failure of their discipline to accumulate proper scientific knowledge that occurred since incorporating the NHT paradigm to their discipline (Loftus, 1996; Rossi, 1997). This discussion has paralleled a reform promoted from medical journal since the mid-1980s (Fidler *et al.*, 2004).

Surprisingly little of this debate has permeated deeply into the daily practice of data analysis in biology and ornithology, despite a substantial number of papers dealing with this topic has accumulated in biological journals during the last decade (Peterman, 1990; Simberloff, 1990; Matloff, 1991; Yoccoz, 1991; Ellison, 1996; Steidl *et al.*, 1997; Johnson, 1999, 2002; Krebs, 2000; Anderson *et al.*, 2001; Colegrave & Ruxton, 2002; Di Stephano, 2003; Eberhardt, 2003; Mogie, 2004; Clark, 2005; Stephens *et al.*, 2005).

Presently there are both radical opponents and defendants of the NHT paradigm, but little grounds for a consensus have been allowed. Here, we shall review criticisms and alternatives to the NHT paradigm, with the goal of highlighting the communalities of both sectors. We have tried to write this review paper in plain language, from the point of view of an average working biologist, not particularly knowledgeable about statistical matters. Following the advice of Krebs (2000), clearly we should spend most of our time thinking about ecological issues; however it is basic that we should have useful procedures for data analysis, which contribute to the proper accumulation of scientific knowledge.

Main Arguments against and for NHST

Against NHST (Table 1)

Argument 1

There are logical and philosophical problems with our methods, such as the fact that in testing and rejecting the null hypothesis researchers focus on a scientifically irrelevant hypothesis and not on the actual hypothesis of biological interest (Schmidt & Hunter, 1997).

Argument 2

The most relevant drawback of NHT comes from the fact that when we test the null hypothesis we are finding the probability that the observed data (i.e. the value of the test statistics obtained), or more, could have arisen provided that the null hypothesis was true (Cohen, 1994). However, what we usually desire is instead the probability that the null hypothesis is correct given our data, the so-called posterior probability.

Argument 3

Our P-value is a probability but a probability conditioned to the truth of the null hypothesis, and we almost always know in advance that the nil hypothesis (i.e. the null hypothesis stating that some parameter equals zero or that some set of parameters are equal) is false to some degree. For example, differences between means may be very small, even close to zero, but not exactly zero as stated by the nil hypothesis. Hence, this implies that if we desire a significant result in ornithology all we have to do is increasing sample size to the necessary extent!

Argument 4

Finally, although we commonly test nil hypotheses in a dichotomous way (i.e. accept/reject) this procedure is dissatisfying because we are more often interested in knowing the precision of our estimate and the size of the effect (i.e. the biological significance or biological

TABLE 1

Summary of main arguments against and for null hypothesis statistical testing and definition of the main technical terms used.
[Principales argumentos a favor o en contra del análisis estadístico de la hipótesis nula. Además se facilita la definición de los principales términos técnicos utilizados.]

Arguments against NHST	**Arguments in favour of NHST**
1. Focus on a scientifically irrelevant hypothesis when testing H_0	1. Adequate when our target is $P(D\|H_0)$
2. We test $P(D\|H_0)$ but we aim for $P(H_0\|D)$	2. Adequate if using specific null hypotheses instead of nil hypotheses
3. We know that H_0 is almost always false beforehand	3. Provides good approximation when sample size is small because we are interested in large effects
4. Inappropriate dichotomous (accept/reject) decision-making procedure	4. Individual studies are useful for meta-analysis

- Point null hypothesis = the traditional null hypothesis stating that a parameter is exactly equal to some individual value.
- Nil hypothesis = the null hypothesis stating that some parameter equals zero or that some set of parameters are equal
- Specific null hypotheses = null hypotheses stating whether an effect is so small that is unimportant biologically
- Meta-analysis = The joint analysis of a set of papers to obtain true conclusions from the point estimates of the parameter, degrees of freedom and estimates of dispersal of individual papers.
- $P(D|H_0)$ = Probability of our data conditioned to the truth of the null hypothesis
- $P(H_0|D)$ = Probability of the null hypothesis given our data (i.e. referred to the value of the statistic employed)

relevance) rather than whether it is precisely zero or not (i.e the statistical significance) (Steiger & Fouladi, 1997). Significance and biological relevance of results are often decoupled in NHT.

In favour of NHST (Table 1)

Argument 1

Provided that we are confident enough that the zero-effect predicted by the nil hypothesis could actually occur (under experimental conditions), it would be perfectly correct to proceed through NHT to find the probability of our data conditioned to the truth of the null hypothesis.

Argument 2

In most instances, when we know beforehand that this prediction (i.e. the zero effect) is false we could make specific predictions about the parameter, based on previous information, and build informative null and alternative hypothesis (Rindskopf, 1997). In that case proceeding through NHT would be correct.

Argument 3

Testing the point null hypothesis can provide a good approximation as to whether the desired population parameter is nearly equal to that hypothesized by H_0, if the sample size is modest (Rindskopf, 1997). This is so because with small sample sizes the power (i.e. the probability of correctly rejecting the null hypothesis) to declare significant a small effect size is very low, whereas the power to detect medium to large effect sizes is be much higher and we are usually interested in medium to large effect sizes in ornithology.

Argument 4
Individual studies based on NHT can be useful in the long run because they can be integrated in meta-analysis if the proper information is available (i.e. point estimates of the parameter, degrees of freedom and estimates of dispersal).

ALTERNATIVES TO NHT

A large number of opponents and defendants of significance testing agree on the need of reforming the paradigm for proper statistical data analysis. A moderate approach would include at least the following major changes:

Alternative 1: Use plausible specific hypotheses with biological meaning

Null hypotheses suggesting that a parameter equals zero are seldom worth testing as ecological models (Krebs, 2000). Hence, we should formulate hypotheses with biological meaning based on previous information regarding our study system (Cohen, 1994). This problem becomes more and more relevant as we accumulate knowledge on a topic because when we continue to formulate nil hypotheses, regardless of our prior knowledge, alternative hypotheses have increased chances of being correct; in such a case the error rate is not measured by α but rather by β (i.e the Type II error rate). Hence the error rate would not be of the usual 5% but much higher (40%-60% as an average in psychology and probably in biology). Specified contrasts should be the rule in ecological studies according to Krebs (2000).

Alternative 2: Use meta-analysis to integrate findings across multiple studies

Only meta-analysis (i.e. the joint study of individual studies) can present credible evidence for or against a scientific hypothesis. Hence, individual contributions should focus on estimation of parameters rather than hypothesis testing (Quinn & Dunham, 1983). In this sense individual papers reporting solely asterisks to represent the level of significance and not providing the value of the statistic and the basic parameter estimates are useless for subsequent integrated analysis (Meehl, 1997). Of course meta-analysis can be misleading if only studies with significant results are published (Johnson, 1999; Quinn & Keough, 2004).

Alternative 3: Report precision and biological relevance in the form of confidence intervals

When testing nil hypotheses we know beforehand that the parameter is always different from zero because there are no two things exactly equal in biology. The questions we actually want to know are the precision of our estimates and the effect size (i.e. the magnitude of the difference between hypothesized and sample values scaled by the within-population standard deviation). An alternative way of reporting all the information we want when we plan a significance test is providing confidence intervals (CIs hereafter) for the difference or the ratio between treatments (Reichardt & Gollob, 1997), a practice already common in many medical journals (Cumming *et al.*, 2004). Confidence intervals are progressively available for many parameters although some of them (*e.g.* chi-square contingency table) have no uniquely defined parameter associated with them (Johnson, 1999). Additionally, they can be derived reliably for most parameters through bootstrap techniques instead of using automated procedures from computer programs (Effron & Tibshirani, 1991; Peres-Neto *et al.*, 2003; Quinn & Keough, 2004). The interpretation of CIs is different for opponents (i.e. subjectivists) and defendants (i.e. frequentists) of NHT from a probabilistic point of view. Frequentists cannot speak of probabilities of finding the population parameter in the confidence interval calculated for a particular study but of levels of confidence (*e.g.* for a 95% CI, if the study were repeated an infinite number of times, 95% of CIs that resulted would contain the true value of the population parameter). On the contrary the former can (i.e. the probability that the value of the parameter lies within the interval is 95% for a Bayesian) because population parameters are considered as random instead of fixed by Bayesians (Pruzek, 1997; Reichardt & Gollob, 1997; Johnson, 1999).

Alternative 4: Use prospective power analysis

Publication in scientific journals is skewed towards papers reporting positive results that are statistically significant (Reichardt & Gollob, 1997). Studies reporting non-significant results are typically rejected for publication often because reviewers and editors presume that small sample sizes are what lead to claiming a lack of effect (i.e. committing a Type II error). The best procedure to prevent this bias is publishing non-significant results of tests with enough power to detect biologically relevant effects (Alchin, 1999). Power () is the probability of correctly rejecting a null hypothesis that is false (Sokal & Rohlf, 1981; Steidl *et al.*, 1997); it is positively linked with the size of α and the sample size, and inversely related to the amount of within-group variability as measured by the effect size (Cohen, 1988; Steidl *et al.*, 1997). Statistical power is considered adequate by convention if its value is 0.80 or above (Cohen, 1988; Steidl *et al.*, 1997), and hence β is set at 0.2 (20%), that is, the Type II error rate is considered four times less costly than the typical α of 0.05 (5%). However, critiques to setting power values arbitrarily have arisen from ecological grounds (Di Stefano, 2003) because in ecological studies Type II errors (i.e. failing to reject a false null hypothesis) are sometimes more costly than Type I errors (Toft & Shea, 1983; Peterman, 1990). Hence a rational solution would be defining the α/β ratio on a case by case basis, both in a priori and post hoc power analyses (Di Stephano, 2003). Software available for power analysis allows performing both a priori and post hoc analysis. Prospective analyses are useful for proper experimental design and post hoc analyses are most appropriate for answering specific questions relating to hypothetical scenarios (Steidl *et al.*, 1997). Interestingly, the width of CIs can provide us with an analogue of power analysis, as larger sample sizes reduce the breadth of CIs and increase power of NHT.

Alternative 5: Use Bayesian methods to provide a measure of evidence of your effect

As previously stated (see Argument 2 againts NHT), the main problem is that we wish to obtain $P(H_0|D)$ but we actually test $P(D|H_0)$. Bayesian statistics solves this problem supporting inferences by use of subjective prior information (Pruzek, 1997). Later collection of empirical data allows construction a conditional distribution $[P(D|H_0)]$ and finally the prior and the conditional "likelihood" are combined to form a posterior distribution $[P(H_0|D)]$ that can be used as a basis for inferences through Bayes' theorem

$$P(H_0|D) = \frac{P(D|H_0)\,P(H_0)}{P(D)}$$

Most importantly Bayesian methods provide numerical intervals (i.e. posterior credible intervals) that represent actual probabilities which make them more natural and useful for decision making (Wolfson *et al.*, 1996; Johnson, 1999). Bayesian methods are especially useful as they provide a measure of the evidence of the effect and its adaptive nature renders meta-analysis unnecessary (Pruzek, 1997; Rindskopf, 1997).

However Bayesian methods have some drawbacks such as the difficulty of obtaining prior probabilities and their complicated implementation until recently (Rozeboom, 1997). However, straightforward Bayesian solutions for most statistical techniques are now available through MCMC (i.e. Markov Chain Monte Carlo) (Beaumont & Rannala, 2004). Owing to this technical advance, Bayesian methods are now employed in a large number of scientific disciplines (Gianola *et al.*, 2002) to solve complex models.

Alternative 6: Use decision theory to support decision making

Hypothesis testing is an inadequate frame for ornithological studies designed to helping decision making on any conservation matter. In those instances statistical decision theory (Westphal & Possingham, 2003) is a better approach because it takes into consideration the cost of alternative actions (Johnson, 1999). Again, in ecological situations Type II errors can be very costly and this fact cannot be over sighted when management is done based on scientific criteria.

Alternative 7: Select models using information criteria

Statistical models build to test specific biological hypothesis can be compared to selecting the best model, i.e. choosing the model which better describes the data with the lowest number of parameters, which is not necessarily the "right" model. Thus, there is a trade off between the number of parameters used in a model and the inherent increase in sampling variance. The problem with NHT regarding model selection is that there is no satisfactory statistical basis for determining which α-level will lead to an appropriate trade-off between over-parameterization and variance, a problem to which the use of information criteria is not vulnerable (see Stephens *et al.*, 2005 for further discussion of the topic). Several information criteria are available to select the best model (see a review of methods in Williams *et al.*, 2001, p. 306), but presently the most commonly used is the Akaike's Information Criterion (AIC) (Anderson *et al.*, 1994; Anderson & Burnham, 1999). This information criterion has the advantage of handling not nested models, which is not possible when using likelihood ratio tests for model selection, and it takes into account the deviance (G^2) (i.e. the difference in fit of the full and reduced models) of the model and its number of parameters. Importantly, a Bayesian equivalent to AIC, the Schwarz Bayesian Information Criterion (BIC), is also available for model selection (Quinn & Keough, 2004).

SUMMARY POINTS FOR A CONSENSUS (Table 2)

It is likely that many long unproductive ornithological debates, such as the determinants of clutch size in birds, with as many papers against and for a given hypothesis, are due to the use, misuse and abuse of NHT, rather than to the intractable nature of the topic. Although radical critics defend that significance testing should be discontinued as a research tool (Schmidt & Hunter, 1997) and that magic alternatives to NHT are not available (Cohen, 1994), we suggest that there are grounds for a possible consensus with a pluralistic approach and that this consensus should be reached as soon as possible. This is an open debate but NHT can probably be maintained as a proper way of statistical analysis of biological data if several points are respected, especially those concerning the use of informative null hypotheses, together with confidence intervals and power analysis. Probably, reporting confidence intervals for the difference between treatments together with P-values of spe-

TABLE 2

Guidelines proposed to change current practices of null hypothesis statistical testing and definition of the main technical terms used.
[Directrices propuestas para el cambio de la práctica actual del análisis estadístico de la hipótesis nula y definición de los principales términos técnicos utilizados.]

Present practice	**Alternative recommended**
1. Use of false nil hypothesis	1. Use of plausible specific null hypothesis with biological meaning
2. Dichotomous (accept/reject) decisions	2. Use of plausible specific null hypothesis with biological meaning
3. Obtain definitive results from individual studies	3. Extend the use of meta-analysis
4. Neglect of the magnitude of the effect (biological significance)	4. Care for biological relevance reporting interval estimates and prospective power analysis to detect specific effect sizes
5. Neglect of non-significant results	5. Assess the validity of non-significant results through power analysis
6. Mistake $P(D\|H_0)$ and $P(H_0\|D)$	6. Use Bayesian statistics to provide a measure of evidence of your data (i.e. unconditional probabilities)
7. Risk of inadequate decisions owing to high Type II error rate	7. Use decision theory for decision-making
8. Inadequate model selection	8. Use information criteria for model selection dealing with multivariate patterns of causality

- Power = the probability of correctly rejecting a null hypothesis that is false.
- Type I error = rejecting the null hypothesis when it is correct.
- Type II error rate = failing to reject the null hypothesis when it is false.
- Type III error rate = statistical significance in the wrong direction.
- Bayesian statistics = school of inferential statistics based on Bayes' theorem. The "likelihood" is combined with prior information to obtain posterior probability distributions of the population parameter.
- Effect size = an index that measures the magnitude of the difference between hypothesized and sample values scaled by the within-population standard deviation.

cific hypotheses is the best option to provide all the information we need (i.e. relevance, direction, precision, power and statistical significance) to determine practical importance and make correct decisions when investigating univariate causality. Likewise, the use of information criteria should be further promoted when multiple causal factors are considered. Importantly, the alternative approach of Bayesian statistics will probably see a large expansion among ornithologists over these coming years, improving their role as scientific advisers of managers in charge of species protection, as it provides reliable measures of evidence of the effects.

ACKNOWLEDGEMENTS.—This work has benefited from critical comments by Robert. E. Ricklefs, Lluis Jover, David Serrano, Agustín Blasco and Mario Díaz. We are most thankful to all of them although any possible mistakes are only attributable to us. Funding of this study is partly due to Conselleria de Territorio y Vivienda (Generalitat Valenciana).

BIBLIOGRAPHY

ALCHIN, D. 1999. Negative results as positive knowledge, and zeroing in on significant problems. *Marine Ecology Progress Series*, 191: 303-305.

ANDERSON, D. R., BURNHAM, K. P. 1999. Understanding information criteria for selection among capture-recapture or ring-recovery models. *Bird Study*, 46: 14-21.

ANDERSON, D. R., BURNHAM, K. P. & WHITE, C. G. 1994. AIC model selection in overdispersed capture-recapture data. *Ecology*, 75: 1780-1793.

ANDERSON, D. R., BURNHAM, K. P., GOULD, W. R. & CHERRY, S. 2001. Concerns about finding effects that are actually spurious. *Wildlife Society Bulletin*, 29: 311-316.

BEAUMONT, M. A. & RANNALA, B. 2004. The Bayesian revolution in genetics. *Nature reviews,* 5: 251-261.

CLARK, J. S. 2005. Why environmental scientists are becoming Bayesians? *Ecology Letters*, 8: 2-14.

COHEN, J. 1994. The earth is round (p<.05). American Psychologist, 49: 997-1003. Reprinted in L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* Lawrence Erlbaum Associates. London.

COHEN, J. 1988. *Statistical power analysis for the behavioural sciences* (2nd ed.). Lawrence Erlbaum Associates, Hillsdale.

COLEGRAVE, N., & RUXTON, G. D. 2002. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology*, 14: 446-450.

DI STEPHANO, J. 2003. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology*, 17: 707-709.

EBERHARDT, L. L. 2003. What should we do about hypothesis testing? *Journal of Wildlife Management*, 67: 241-247.

EFFRON, B. & TIBSHIRANI, R. 1991. Statistical data analysis in the computer age. *Science*, 253: 390-395.

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

FIDLER, F., THOMASON, N., CUMMING, G., FINCH, S. & LEEMAN, J. 2004. Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, 15: 119-126.

FISHER, R. A. 1951. *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd. (Original work published in 1925).

GIANOLA, D., REKAYA, R., ROSA, G. J. M. & SANCHES, A. 2002. Communication Nº 17-02. 7th World Congress on Genetics Applied to Livestock Production, August 19-23, 2002, Montpellier, France.

JAMES, F. C. & MCCULLOGH, C. E. 1990. Multivariate analysis in ecology and systematics: Panacea or Pandora's box? *Annual Review of Ecological Systematics*, 21: 129-166.

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.

KREBS, C. J. 2000. Hypothesis testing in ecology. In, L. Boitani & Fuller, T. K. (Eds): *Research techniques in animal ecology: controversies and consequences*, pp. 1-14. Columbia University Press, New York.

LOFTUS, G. R. 1996. Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5: 161-171.

MATLOFF, N. S. 1991. Statistical hypothesis testing: problems and alternatives. *Environmental Entomology*, 20: 1246-1250.

MEEHL, P. E. 1997. The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In, L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* pp. 393-425. Lawrence Erlbaum Associates. London.

MOGIE, M. 2004. In support of null hypothesis significance testing. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 271: S82-S84.

MULAIK, S. A., RAJU, N. S. & HARSHMAN, R. A. 1997. There is a time and a place for significance testing. In, L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* pp. 65-115. Lawrence Erlbaum Associates. London.

OAKES, M. 1986. *Statistical inference: a commentary for the social and behavioural sciences*. Wiley, New York.

PERES-NETO, P. R., JACKSON, D. A. & OMERS, K. M. S. 2003. Living meaningful interpretation to ordination axes assessing loading significance in principal component analysis. *Ecology*, 84: 2347-2363.

PETERMAN, R. M. 1990. Statistical power can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences*, 47: 2-15.

PRUZEK, R. M. 1997. An introduction to Bayesian inference and its applications. In. L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* pp. 287-318. Lawrence Erlbaum Associates. London.

QUINN, J. F. & DUNHAM, E. 1983. On hypothesis testing in ecology and evolution. *American Naturalist*, 122: 602-617.

QUINN, G. P. & KEOUGH, M. J. 2004. *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge.

REICHARDT, C. S. & GOLLOB, H. F.1997. When confidence intervals should be used instead of statistical tests, and vice versa. In, L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* pp. 259-284. Lawrence Erlbaum Associates. London.

ROSSI, J. S. 1997. A case study in the failure of psychology as a cumulative science: the spontaneous recovery of verbal learning. In, L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* pp. 175-197. Lawrence Erlbaum Associates. London.

ROZEBOOM, W. W. 1997. Good science is abductive, not hypothetico-deductive. In, L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* pp. 335-392. Lawrence Erlbaum Associates. London.

SALSBURG, D. S. 1985. The religion of statistics as practiced in medical journals. *The American Statistician*, 39: 220-223.

SCHMIDT, F. L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1: 115-129.

SCHMIDT, F. L. & HUNTER, J. E. 1997. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In, L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* pp. 37-64. Lawrence Erlbaum Associates. London.

SIMBERLOFF, D. 1990. Hypothesis, errors, and statistical assumptions. *Herpetologica*, 46: 351-357.

STEIDL, R. J., HAYES, J. P. & SCHAUBER, E. 1997. Statistical power analysis in wildlife research. *Journal of Wildlife Management*, 61: 270-279.

STEIGER, J. H. & FOULADI, R. T. 1997. Noncentrality interval estimation and the evaluation of statistical models. In, L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds): *What if there were no significance tests?* pp. 221-258. Lawrence Erlbaum Associates. London.

STEPHENS, P. A., BUSKIRK, S. W., HAYWARD, G. D. & MARTÍNEZ DEL RÍO, C. 2005. Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, 42: 4-12.

TOFT, C. A. & SHEA, P. J. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. *American Naturalist*, 122: 618-625.

WESTPHAL, M. I. & POSSINGHAM, H. P. 2003. Applying a decision-theory framework to landscape planning for biodiversity: Follow-up to Waston *et al. Conservation Biology*, 17: 327-329.

WILLIAMS, B. K., CONROY, M. J. & NICHOLS, J. D. 2001. *Analysis and Management of Animal Populations*. Academic Press, San Diego.

WOLFSON, L. J., KADANE, J. B. & SMALL, M. J. 1996. Bayesian environmental policy decisions: two case studies. *Ecological Applications*, 6: 1056-1066.

YOCCOZ, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, 72: 106-111.

[Recibido: 09-02-05]
[Aceptado: 13-09-05]