REVISIONES

Data analysis in forest sciences: why do we continue using null hypothesis significance tests?

Análisis de datos en ciencias forestales: ¿por qué continuar usando pruebas de significancia estadística?

Sergio A Estay [a]*, Paulette I Naulin [b,c]

*Corresponding author: [a] Pontificia Universidad Católica de Chile, Center for Advanced Studies in Ecology and Biodiversity,
casilla 114-D, Santiago, CP 6513677, Chile, tel.: 562 3542609, fax: 562 3542621, sestay@bio.puc.cl
[b] Universidad de Chile, Departamento de Ciencias Ecológicas, Santiago, Chile.
[c] Universidad de Chile, Departamento de Silvicultura y Conservación de la Naturaleza, Santiago, Chile.

SUMMARY

Statistical methods are indispensable for scientific research. In forest sciences, the use of null hypothesis significance tests (NHSTs) has been the rule of thumb to judge hypotheses or associations among variables, in spite of the multiple problems of these techniques and the several criticisms published for many years in other scientific areas. In this review, the origin of current techniques, their most important problems, and some alternatives that are known to most forest researchers are shown. Persistence in using NHSTs, instead of better statistical methods or without adequate complements, could render our work inefficient and risky. Reasons for the permanence of NHSTs in forest sciences are discussed.

Key words: NHST, p-values, statistical significance, information criteria, ANOVA.

RESUMEN

Los métodos estadísticos son una parte indispensable del quehacer científico. En las ciencias forestales el uso de pruebas de significancia estadística ha sido la regla predominante para juzgar hipótesis o asociaciones entre variables a pesar de sus múltiples problemas y las diversas críticas publicadas por muchos años en otras áreas de la ciencia. En esta revisión se muestra el origen de las actuales metodologías, sus principales problemas y se presentan algunas opciones al alcance de la mayor parte de los investigadores. De continuar usando estas técnicas en lugar de métodos estadísticos correctos o sin el adecuado complemento, el trabajo podría tornarse ineficiente y riesgoso, en especial, dadas las importantes decisiones que en materia medioambiental corresponde tomar. Razones para la permanencia de estas pruebas en las ciencias forestales son discutidas.

Palabras clave: prueba de significancia estadística, valor de p, hipótesis nula, criterios de información, ANDEVA.

INTRODUCTION

The work of foresters, as professionals directly related to nature, is focused on addressing the causes of natural phenomena. During this process, a set of methods learned during our undergraduate studies, and which usually seem to be beyond question, are used. Among these are null hypothesis significance tests, NHSTs (ANOVA, student t test, Chi square, among others). NHSTs have been criticized almost since their very origin (Berkson 1938, 1942). However, these criticisms have not been acknowledged enough in the forestry area. This review shows some of the most important problems and limitations of these methods, their scope, possible solutions and the reason of the persistent use of these methods. This work does not intend to be an exhaustive review or an original idea, but it is necessary to generate a discussion given the magnitude and importance of decisions related to natural resources that are supported by data analyses.

WHAT ARE NHSTs AND WHAT IS THEIR ORIGIN?

Current NHSTs are the outcome of the combination of the foundational works of Fisher (1925, 1935), and Neyman and Pearson (1928ab) (Spielman 1974, Goodman 1993, Gill 1999, Anderson *et al.* 2000).

Fisher (1925, 1935) proposes that a statistical hypothesis H_0 must be defined, which has a known distribution for the statistic T. First, the value and probability (the p-value) of T should be calculated from the data. Finally, H_0 must be rejected if the significance level is small enough. On the other hand, Neyman and Pearson (1928ab) proposed to evaluate two complementary hypotheses H_0 and H_1. First, a priori values for α and β (or Type I and Type II errors) are defined. Given this, Neyman and Pearson define the power of the test, 1- β, which is the probability of rejecting a false null hypothesis. Once the hypotheses are defined, the researcher must use the test with the highest power. If T is higher than a critical value α, then reject H_0 and accept H_1; otherwise accept H_0.

The synthesis of these two approaches leads us to the current NHSTs: two complementary hypotheses are evaluated with a predefined critical value, but calculating a p-value and forgetting to search for the test with the highest power. Many of the problems of the current methodology come from the fusion of these two approaches (Goodman 1993, Gill 1999). However, before continuing to examine the criticisms of these methods, it is convenient to look at the current status of their use in forest sciences.

THE USE OF NHSTs IN FOREST SCIENCES

In order to have a preliminary notion of the influence of NHSTs in forest sciences, we carried out a review of two of the most influential journals in forest science in the world, namely "Forestry" and "Forest Ecology and Management". All issues between January 2009 and July 2010, 82(1-5) - 83(1-3), in Forestry and the full year 2009, 259(1-12), in Forest Ecology and Management were reviewed. All articles that contained some statistical evaluation together with p-values in their results were identified. We did not make any kind of judgment about the relevance of those p-values in the conclusions of each study. The latter would have demanded a detailed reading of each paper, which is out of the scope of this review. In Forestry, from 69 regular articles, 42 (61 %) had at least one p-value in their results. In Forest Ecology and Management, 124 of 180 papers had at least one p-value (69 %).

It is clear that the use of NHSTs is widespread in forest science. The previous review considers only two journals, but it would be very interesting to know what proportion of professional reports in private companies and public services uses this methodology and how relevant those results are for decision-making.

PROBLEMS WITH NHSTs

What does the p-value mean? There are at least three common interpretations of the p-value often repeated in the literature (Carver 1978). The first is that the p-value is the degree of replicability of the result, being 1-p the probability that a replicate of the study also yields a significant result (Carver 1978, Johnson 1999, Nickerson 2000, Kline 2004). This is called the "replication fallacy" (Falk and Greenbaum 1995, Gill 1999, Kline 2004). The second interpretation is that the p-value is the probability of obtaining the same outcome by chance or the probability that the outcome could only be the result of the sample selection process. This is called the "odds-against-chance" fallacy (Carver 1978, Falk and Greenbaum 1995, Johnson 1999, Kline 2004). Finally, the p-value is interpreted as the probability of H_0 being true, that is $Pr(H_0|Data)$ which is called the "inverse probability" fallacy (Carver 1978, Cohen 1994, Johnson 1999, Nickerson 2000, Kline 2004). Of course, the question is: if all of these interpretations are incorrect, then what does the p-value mean? Figure 1 shows a probability density function with its mean μ. The dark area represents the α value.

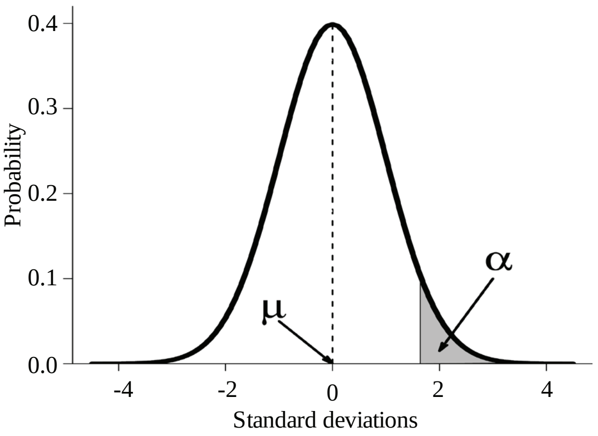

Figure 1. Example of a normal probability density function. μ represents the mean value and the dark area represents the α value.

Ejemplo de una función de densidad de probabilidad normal. μ representa la media y el área sombreada representa el valor de α.

When a null hypothesis, H_0, is defined, then the area under the curve between the calculated T value (T_c) and ∞ and/or -∞ is the p-value, $Pr(T \geq T_c \mid H_0)$. The p-value is the probability of obtaining our calculated T value or a more extreme one, given that the null hypothesis is true (Gill 1999, Johnson 1999). With the correct definition, the problems of the previous interpretations may be detected. In the first case, it is clear that the p-value does not represent the confidence in the test. Since the p-value is calculated from our set of data, $Pr(Data \mid H_0)$, the p-value does not say anything about what the distribution of the T statistic is, given multiple sets of data (Gill 1999). The second and third interpretations have the same problem: they assume that the p-value is the probability of H_0 given the data, $Pr(H_0 \mid Data)$. Nevertheless, the p-value is calculated assuming H_0 is true, $Pr(Data \mid H_0)$. Both probabilities are not the same (Carver 1978, Cohen 1994, Falk and Greenbaum 1995, Gill 1999, Johnson 1999, Nickerson 2000). From the Bayes theorem:

$$Pr(H_0 \mid Data) = \frac{Pr(Data \mid H_0)Pr(H_0)}{Pr(Data)} \qquad [1]$$

At this point the reader may feel that these are just problems of interpretation, and that an informed researcher can obtain the real meaning from NHSTs. However, the next objections show that NHSTs really contribute little to the discovery of the underlying mechanisms of any phenomenon.

Null hypotheses and their relevance. When NHSTs are used, it is assumed that the population parameter exists and is fixed at some value; this is the null hypothesis. But, how plausible is it that samples taken from different populations (or treatments) have exactly the same parameter value that the null hypothesis suggests? For example, what is the probability that variables like DBH, density, above ground biomass or any others have the same value in two or more stands? The null hypotheses are known to be false before beginning the analysis (Berkson 1938, Cohen 1994, Johnson 1999, Anderson *et al.* 2000, Nickerson 2000). The only information that a NHST gives us is whether the sample size was large enough to detect the difference (Yoccoz 1991). For example, if the values of sample sizes, means, and variances for two samples are N_1, N_2, $\bar{X}_1$, $\bar{X}_2$, S_1^2 and S_2^2, then the t-student statistic would be,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} \qquad [2]$$

When sample sizes increase, means and variances will become stabilized around their real values; but even though the difference between means is small, if N_1 and N_2 increase, then *t* increases, and the p-value decreases until it becomes smaller than 0.05. The p-value is arbitrary; any association will be significant if the sample size increases enough.

In forest sciences it is possible to find several examples where different populations or stands of some organisms are compared using NHSTs, even between localities. As a researcher or professional in charge of a study, the question is not if A and B are different, as Tukey said in 1991, but what the direction and magnitude of the difference is. And the answer will never be obtained from a NHST (Anscombe 1956, Tukey 1991, Frick 1996, Martínez-Abraín 2007).

Statistical, theoretical and/or practical significance. One of the most relevant and less emphasized aspects of any research is the - theoretical or practical- significance of the results. Let us assume that a statistically significant difference in an experiment is found, and it is concluded that treatment X has an effect on, for example, the growth of seedlings of species Y in a nursery. However, it is possible to see that the increment due to treatment X is on average 0.5 cm, which for species Y, and its culture, is irrelevant. This takes us back to the previous point: despite the statistical significance, the most important thing is the direction and magnitude of the effect (Yoccoz 1991). It is important to remember that it is the theoretical framework in which the problem is placed where the real significance of the results will be found (Yoccoz 1991).

Additional criticisms to NHSTs. One of the more serious criticisms towards NHSTs is their logical inconsistency. The logical base of the tests lies in the syllogism called *modus tollendo tollens* or indirect reasoning (Berkson 1942, Falk and Greenbaum 1995, Gill 1999, Nickerson 2000). This syllogism works in the following way:

If A then B
B is false
Then A is false

In terms of our test, the latter is equivalent to:

If H_0 is true then the data must follow the pattern X
The data do not follow the pattern X
Therefore H_0 is false

This syllogism is valid in the case of categorical premises, but it is invalid with probabilistic premises (Cohen 1994, Falk and Greenbaum 1995, Nickerson 2000). Notice that:

If H_0 is true then probably $P > 0.05$
$P < 0.05$
Therefore H_0 is probably false

There are no logical reasons to doubt the veracity of H_0 given that a rare event has occurred (Spielman 1974). Furthermore, it is important to notice that H_0 makes reference to an exact value (usually 0), but H_1 represents all other infinite possibilities. The logical inconsistency of NHSTs is a fact acknowledged by their followers, who recognize some usefulness despite these inconsistencies (Chow 1988, Hagen 1997, 1998). As Falk (1998) pointed out, defending a logically invalid argument could make some sense if experience indicates some valid relationship. However, when the test has an inconsistency that might, for example, make us reject a null hypothesis that a posteriori has a reasonable or high probability to be true, it becomes unacceptable (Berger and Sellke 1987, Falk 1998).

ALTERNATIVE STATISTICAL TECHNIQUES FOR A BETTER DATA ANALYSIS

Forestry literature is plentiful in articles about new methods or new statistical approaches other than NHSTs, but usually these new approaches are mixed with p-values and other characteristics of NHSTs without recognizing the deep differences among them. In order to facilitate the decision about what the best analytical approach for a specific problem is; in this section we present some alternatives, their scope, requirements and advantages/disadvantages of their use.

Suggested complements for NHSTs. If the researcher feels comfortable about NHSTs despite the previous arguments, and the use of a new technique is not possible, then we strongly recommend complementing them with some analyses such as:

• Confidence intervals: they provide information about the effect size and the uncertainty of the estimation (Johnson 1999, Walshe *et al.* 2007), which avoids the dichotomous thinking associated to NHSTs. This makes them an essential component for any data analysis. Gardner and Altman (1986) pointed out that one of their more important advantages is that they make the understanding of the results easier, since confidence intervals show results in the same scale in which the measurement was obtained. Moreover, confidence intervals provide the necessary information for future meta-analyses (Fidler *et al.* 2006). However, it is important to correctly interpret the concept of confidence. In the case of traditional statistics, interpretations are based on the frequency of a specific result in many replicates (for this reason it is called "frequentist" in some textbooks), and where the true value of the parameter is fixed but unknown. So the probability that a particular interval contains the true value of the parameter is one or zero. In other words, this kind of interval does not say anything about the confidence of our particular result. The correct interpretation of a confidence interval of, for example, 95 % is that if the study were repeated many times, in 95 % of cases the interval would contain the real value of the parameter (Johnson 1999). A guide to confidence intervals use are Altman *et al.* (2000) and a series of articles by Geoff Cumming (Cumming and Finch 2005, Cumming 2007, 2009).

• Range null hypotheses: Nickerson (2000) suggested that the use of point NHSTs should be changed to range NHSTs, where all the values in a range are considered to belong to the null hypothesis. In practice this is what most researchers believe that NHSTs do: there is a range of values around the value predicted by the null hypothesis whose difference could be considered negligible (Greenwald 1975). Among the advantages of this principle is the obligation of defining the direction and magnitude of the expected change in the focal parameter given the treatments, and with this information, incorporate the practical significance in the analysis of statistical significance. Range NHSTs are similar to the "good enough principle" proposed by Serlin and Lapsey (1985). This principle demands to define a priori what a "good difference" is, in order to support the working hypothesis.

Other approaches. NHSTs are not the only methodology or philosophy that can be used to evaluate hypotheses in science. Other approaches may take a different theoretical framework, such as information theory, or a complete different concept of probability, such as Bayesian methods. In the next paragraphs some of these approaches are examined.

• Maximum likelihood and information criteria: as early as 1890, Thomas C Chamberlain advocated the use of multiple working hypotheses instead of a single hypothesis. Using this approach, there is not a null hypothesis but several hypotheses, usually sustained in previous studies, which are evaluated in terms of their relative support in the data. This approach is called model selection or multimodel inference in the specialized literature.

Mathematical models are a very powerful tool that allows us to avoid the ambiguity of language and, through their use, make an exact description of our working hypothesis. If each hypothesis is represented by a mathematical model, the next step is to rank them according to their support in the data. It is in this context that the principle of maximum likelihood has an essential role. The principle (of which Fisher is considered the founder) states that among all the alternative hypotheses to explain a phenomenon, we must select the one that maximizes the probability of our data.

In this approach the interest is in the likelihood of each hypothesis given the data, and this is proportional to the probability of the data given each hypothesis:

$$L(H_0 \mid Data) \propto Pr(Data \mid H_0) \qquad [3]$$

In general, we can compare the likelihood of several hypotheses through the likelihood ratio:

$$\frac{L(H_0 \mid Data)}{L(H_1 \mid Data)} \qquad [4]$$

The strength of the evidence in favor of H_0 or H_1 depends on whether the value is greater or lesser than one (Goodman and Royal 1988, Goodman 1993). However, the use of a likelihood ratio test is restricted to nested models (Hilborn and Mangel 1997). To compare models that are not nested it is necessary to appeal to another type of analysis, in specific information criteria such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC or Schwarz information criterion). Next, we review the AIC and its relationship with Log likelihood.

Akaike information criterion relates the concepts of Kullback-Leibler information and maximum likelihood (Anderson *et al.* 2000). Kullback-Leibler information is a concept from physics to measure the difference between reality and the model with which we try to approximate it (see Burnham and Anderson 2004, for a detailed description of this idea). Hirotsugu Akaike (1974) noticed that the Log likelihood of a model is an estimator of the Kullback-Leibler information, but biased. Nevertheless, he also realized that the bias was equal to the number of parameters of the model. Therefore, he defined his information criterion as:

$$AIC = 2\,\mathrm{K} - 2\mathrm{Ln}(L(H_i \mid Data)) \qquad [5]$$

Where K is the number of parameters and $L(H_i \mid Data)$

is the likelihood of the model *i* given the data. Since the objective is to minimize the loss of information, the model with the smallest AIC has the highest support in the data.

For a complete review of these and other criteria see Anderson *et al.* (2000), Burnham and Anderson (2002) and Burnham and Anderson (2004). Allometric equations to estimate biomass or volume in forest plantations or native stands, where we have several potential models to explain the phenomenon, are an interesting field to use these criteria.

- Bayesian statistics. Among the alternative philosophies to frequentist methods, Bayesian statistics have shown the fastest development in the last decades. As it was previously reviewed, the Bayes theorem truly allows for evaluation of the probability of a hypothesis given the data.

$$Pr(H_0 \mid Data) = \frac{Pr(Data \mid H_0)Pr(H_0)}{Pr(Data)} \qquad [6]$$

Also, sometimes Pr(H_0) is unknown; therefore some a priori idea about its distribution is needed. This aspect of Bayesian inference has been considered its main disadvantage, because the a priori estimation of Pr(H_0) could be very subjective. However, this disadvantage could at the same time be a major advantage. The estimation does not arise necessarily from nothing. Usually it is supported by a theoretical and empirical framework. For this reason, this approach is related in a better way to the logical bases of knowledge accumulation (Ellison 1996, Hobbs and Hilborn 2006). The previous studies allow us to have some degree of belief in our ideas, hypotheses or theories (in Bayesian statistics the probability is related to the level of belief in something, not in the long term frequency of expected outcomes).When we get a new data set, we update our degree of belief using preceding knowledge and the one obtained with the new study (Hobbs and Hilborn 2006).

Almost every method developed in the framework of frequentist statistics has its counterpart in Bayesian methods, usually supported in Markov Chain Monte Carlo methods. It is not the objective of this paper to make a detailed review of this topic. For a general introduction we recommend the books by Carlin and Louis (2000) and Bolstad (2004), and three excellent books about the application of Bayesian methods to ecology are McCarthy (2005), Kéry (2010) and Link and Barker (2010).

Conclusions or decisions. Especially for applied research, the analyst needs to evaluate several possible decisions and their consequences rather than hypotheses. It is in this context that the difference between conclusions and decisions becomes important, as Tukey (1960) explained. Conclusions are a final proposition reached after the evaluation of the evidence. For example, after an experiment about the influence of nitrogen on plant growth, we can "conclude" that a higher concentration of nitrogen in the soil accelerates plant growth or increases plant productivity. On the other hand, a decision is the act of choosing among several alternatives, considering the advantages and disadvantages of each one, in order to achieve a specific objective. If, for example, to control an insect pest there are several control strategies or methods such as silvicultural, chemical or biological control, then we can "decide", after considering advantages and disadvantages, to control the pest through a mix of silvicultural and biological control strategies. It is in cases similar to the later example when decision theory emerges as an efficient method to face these problems.

- Decision theory and risk analysis: decision theory is focused not only on the probability of error but also on the cost function of these errors (Johnson 1999). Among the methodologies that use decision theory, risk analysis has many applications in prevention of biological invasions, pollution, species extinction, forest fires, and disease prevention, among others. Decision theory could be used in many situations in forest science, *e.g.* in harvesting decisions under different scenarios such as market fluctuations or regular pest outbreaks. Molak (1997) presents an excellent summary of the methods and applications of risk analysis, especially some chapters related to natural resources management. Hannson (1994) reviews the basic aspects of decision theory.

WHY DO WE CONTINUE USING NHSTs?

There has been considerable debate about the usefulness of NHSTs in some scientific areas during the last 40 years. In medical sciences for example, the editorial boards of many journals recommend the use of techniques different from NHSTs, *i.e.* the American Medical Association or the American Psychological Association (Cumming and Finch 2005, Fidler *et al.* 2006, ICMJE 2006). In spite of this, there are still areas where this debate is absent. In forest sciences the discussion has not been relevant. Few recent papers about the topic have been published in journals focusing on these areas. Only in 2007 was a discussion on this topic published in an ecological journal (Stephens *et al.* 2007, Gibbons *et al.* 2007, Martínez del Río *et al.* 2007).

Much of the persistent use of NHSTs is probably due to the influence of the hypothetical-deductive method in scientific thinking (Hilborn and Mangel 1997). Two critics can be made in this point: first, this method has, in general, many weaknesses, in particular in sciences where maintaining the *ceterius paribus* condition is difficult or impossible (Hilborn and Mangel 1997). This is the case of forest sciences, where the researcher defines what fraction of the ecosystem, or eco-fraction, will be studied and generates the hypotheses that will be evaluated or validated through NHSTs or another method. In these areas it is more important to measure the relative contribution of each potential cause than to evaluate a particular hypothesis independently (Quinn and Dunham 1983). Second, even if this philosophical framework is adequate for a particular situation (*e.g.* quality control), NHSTs do not meet the requirements of popperian falsifiability. According to Popper (1963), the most important demarcation criteria were that theories should be subjected to risky test. However, to test nil null hypotheses, which are almost always false, is not, in any case, risky (Fidler 2005, and references therein).

Another characteristic associated to the hypothetical-deductive method that exerts a great influence in the persistence of NHSTs is dichotomous decision making. In most statistical methods courses, the emphasis is in whether or not the null hypothesis is rejected, as the classical hypothetical-deductive method prescribes. We really believe that if teaching emphasis changed from rejection (or not) of null hypothesis to the estimation of parameters and evaluation of uncertainty, then the understanding and appropriate use of classical and new statistical methods would increase among practitioners. In this direction, Fidler *et al.* (2004) advocate for a reform in teaching methods and editorial policies in ecological sciences in order to end the tyranny of p-values.

Rozeboom (1960) suggested that researchers may be just users of methods without critical attitudes towards them. We agree with this statement only in part. Probably this is not the real situation, but it is the greatest risk. As professionals trained in the management of natural resources, foresters make decisions whose impacts are perceived by the whole society. It is important to remember that, despite the fact that it is impossible to be experts in every topic, it is necessary to take responsibility for each result or conclusion and to revise, lucidly and critically, the methods used so to avoid mistakes, without having to paralyze our work.

CONCLUSIONS

Through this review the extensive use of NHSTs in forest science is confirmed, despite the multiple flaws in its logical structure that several authors have demonstrated. The prevalence of NHSTs could have serious consequences for an efficient advance of forest science and industry. To avoid this, some available alternatives were presented, together with their scope and examples.

Perhaps many of the readers will doubt the arguments given here, since they probably think that if these methods were wrong, then science and industry would not make progress as they do. Nevertheless, this reasoning is incomplete. The proficiency and capacity of scientists and professionals have managed to overcome these obstacles. The true question is if progress has been as fast as it could have been. The conclusion of this review seems to be that forest sciences have advanced in spite of the method.

ACKNOWLEDGEMENTS

The authors thank Rosa Scherson, Eli Trowbridge, Mauricio Lima, Daniela Dominguez, Leah Dudley, Sergio Alvarado and two anonymous reviewers for very useful discussions and comments. SE acknowledges the financial support of FONDAP/FONDECYT grant 1501–0001 (Program 2). PN acknowledges the financial support of a CONICYT doctoral fellowship and Laboratorio de Biología Vegetal, Departamento de S y CN, Universidad de Chile.

REFERENCES

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716-722.

Anderson DR, KP Burnham, WL Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management* 64: 912-923.

Anscombe FJ. 1956. Discussion on Dr. David's and Dr. Johnson's Paper. *Journal of the Royal Statistical Society, Series B* 18: 24-27.

Berger JO, T Sellke. 1987. Testing a point null hypothesis: the irreconcilability of p-values and evidence. *Journal of the American Statistical Association* 82:112-122.

Berkson J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* 33: 526-536.

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

Bolstad WM. 2004. Introduction to bayesian statistics. New Jersey, USA. John Wiley. 376 p.

Burnham KP, DR Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods Research* 33: 261 – 304.

Carlin B, T Louis. 2000. Bayes and empirical bayes methods for data analysis. Boca Raton, USA. Chapman & Hall/CRC. 440 p.

Carver RP. 1978. The case against statistical significance testing. *Harvard Educational Review* 48: 378-399.

Chamberlin TC. 1890. The method of multiple working hypotheses: *Science* (old series) 15: 92-96.

Chow SL. 1988. Significance test or effect size? *Psychological Bulletin* 103(1):105-110.

Cohen J. 1994. The earth is round (p < 0.05). *American Psychologist* 49: 997-1003.

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

Cumming G. 2009. Inference by eye: reading the overlap of independent confidence intervals. *Statistics in Medicine* 28: 205-220.

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

Ellison AM. 1996. An introduction to bayesian inference for ecological research and environmental decision-making. *Ecological Applications* 6: 1036-1046.

Falk R. 1998. In criticism of the null hypothesis statistical test. *American Psychologist* 53:798-799.

Falk R, CW Greenbaum. 1995. Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory & Psychology* 5(1): 75-98.

Fidler F. 2005. From statistical significance to effect estimation: statistical reform in psychology, medicine and ecology. PhD Thesis History and Philosophy of Science. Melbourne, Australia. Department of History and Philosophy of Science. University of Melbourne. 275 p.

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

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

Fisher RA. 1925. Statistical methods for research workers. Edinburgh, UK. Oliver and Boyd. 244 p.

Fisher RA. 1935. The design of experiments. Edinburgh, UK. Oliver and Boyd. 350 p.

Frick RW. 1996. The appropriate use of null hypothesis testing. *Psychological Methods* 1: 379-390.

Gardner MJ, DG Altman.1986. Confidence intervals rather than p-values: estimation rather than hypothesis testing. *British Medical Journal* 292:746-750.

Gibbons JM, N Crout, J Healey. 2007. What role should null-hypothesis significance tests have in statistical education and hypothesis falsification? *Trends in Ecology and Evolution* 22: 445–446.

Gill J. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52:647-674.

Goodman SN. 1993. P-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* 137:485-496.

Hagen RL. 1997. In praise of the null hypothesis statistical test. *American Psychologist* 52: 15-24.

Hagen RL. 1998. A further look at wrong reasons to abandon statistical testing. *American Psychologist* 53: 801-803.

Hansson S. 2005. Decision theory: a brief introduction. Estocolmo, Suecia. Uppsala University. 94 p.

Hilborn R, M Mangel. 1997. The ecological detective: confronting models with data. New Jersey, USA. Princeton University Press. 315 p.

Hobbs NT, R Hilborn. 2006. Alternatives to statistical hypothesis testing in ecology: A guide to self teaching. *Ecological Applications* 16: 5-19.

ICMJE (International Committee of Medical Journals Editors). 2006. Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. ICMJE. Accessed Ago. 17, 2010. Available at http://www.icmje.org/urm_main.html

Johnson DH.1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763-772.

Kéry M. 2010. Introduction to Winbugs for ecologists. San Diego, USA. Academic Press. 302 p.

Kline, RB. 2004. Beyond significance testing. Washington, DC, USA. American Psychological Association. 325 p.

Link WA, RJ Barker. 2010. Bayesian inference with ecological applications. San Diego, USA. Academic Press. 339 p.

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

Martínez del Río C, S. Buskirk, PA Stephens. 2007. Response to Gibbons *et al.*: Null-hypothesis significance tests in education and inference. *Trends in Ecology and Evolution* 22: 446.

McCarthy MA. 2007. Bayesian methods for ecology. Cambridge, UK. Cambridge University Press. 296 p.

Molak V. 1996. Fundamentals of risk analysis and risk management. Florida, USA. CRC Press. 496 p.

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

Neyman J, E Pearson. 1928b. On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika* 20:263-294.

Nickerson RS. 2000. Null hypothesis significance testing: a review of an old and continuing controversy *Psychological Methods* 5:241-301.

Popper K. 1963. Conjectures and refutations: the growth of scientific knowledge. New York, USA. Routledge. 582 p.

Quinn JF, AE Dunham. 1983. On hypothesis testing in ecology and evolution. *The American Naturalist* 122: 602-617.

Rozeboom WW. 1960. The fallacy of the null hypothesis significance test. *Psychological Bulletin* 57: 416-428.

Serlin RC, DK Lapsey.1985. Rationality in psychological research: the good-enough principle. *American Psychologist* 40:73-83.

Sthepens PA, S Buskirk, C Martínez del Río. 2006. Inference in ecology and evolution. *Trends in Ecology and Evolution* 22(4): 192-197.

Spielman S. 1974. The logic of tests of significance. *Philosophy of Science* 41: 211–226.

Tukey JW. 1960. Conclusions vs. decisions. *Technometrics* 2: 423-433.

Tukey JW. 1991. The philosophy of multiple comparisons. *Statistical Science* 6: 100-116.

Walshe T, B Wintle, F Fidler, M Burgman. 2007. Use of confidence intervals to demonstrate performance against forest management standards. *Forest Ecology and Management* 247: 237-245.

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

Recibido: 09.08.10
Aceptado: 27.12.10