Revista Chilena de Ornitología 28(1): 1-12
Unión de Ornitólogos de Chile 2022

TOWARDS THE WISE USE OF STATISTICAL SIGNIFICANCE IN THE REVISTA CHILENA DE ORNITOLOGÍA

Hacia el uso sensato de la significancia estadística en la Revista Chilena de Ornitología

RICARDO A. FIGUEROA[1,2] & SERGIO ALVARADO[2,3,4,5]

[1]Editor Jefe
[2]Unión de Ornitólogos de Chile (AvesChile).
Correspondencia: revistachilenaornitologia@aveschile.cl
[3]Editor Estadístico
[4]Programa de Bioestadística, Instituto de Salud Poblacional, Facultad de Medicina, Universidad de Chile.
[5]Laboratorio de Ecología de Vida Silvestre (LEVS), Facultad de Ciencias Forestales y de la Conservación de la Naturaleza, Universidad de Chile.
Correspondencia: salvarado@uchile.cl

Editors' note. - In the previous issue, we published this editorial with Spanish-speaking authors and readers in mind because they constitute most of our audience. However, several non-Spanish-speaking authors have published the results of their studies in our journal, and others may intend to do so. Therefore, for our message also to reach them, we decided to publish this editorial in English. Beyond languages, we hope that the central message of our editorial will stimulate reflection and critical thinking among our audience.

Nota de los editores. - En el número anterior publicamos esta editorial pensando en los autores y lectores hispanoparlantes debido a que constituyen la mayor parte de nuestro público. Sin embargo, varios autores no hispanoparlantes han publicado los resultados de sus estudios en nuestra revista, y es posible que otros tengan la intención de hacerlo. Así, para que nuestro mensaje llegue también a ellos, hemos decidido publicar esta editorial en inglés. Más allá de los idiomas, esperamos que el mensaje central de esta editorial estimule la reflexión y el pensamiento crítico de toda nuestra audiencia.

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. Fisher (1935).

Statistics often get a bad name because it is so easy to misuse them, unintentionally or not. Sylvan Barnet & Hugo Bedau (1993).

... the inferential capabilities of statistics (whether Bayesian, frequentist, or other) are still profoundly limited. Greenland & Poole (2012).

Statistics plays a pivotal role in the scientific process and progress (Hampel 1997, Scheiner 2001). Like other groups of researchers, ornithologists use statistical tests or models to infer patterns or processes and to establish generalizations and predictions from the results of our research (Buckland 1982, Fowler & Cohen 1996, Underhill 1999). This process, known as statistical inference, allows us to deepen our empirical and theoretical knowledge about how birds respond to ecological and environmental changes (James & McCulloch 1985, North & Byron 1985, North 1994, Underhill 1999). However, the use of inferential statistics requires sound theoretical and methodological justification. Therefore, the specific evaluation of statistical analyses is one of the relevant phases of the editorial process in any scientific journal (Parker *et al*. 2018).

In our experience, defective statistical analyses are among the main weaknesses of many manuscripts submitted to ornithological journals. The main shortcomings are that they are incorrect, unjustified, overestimated, or even unnecessary. Moreover, many reviewers familiar with the central topic of a manuscript are not necessarily knowledgeable about statistical theories. Therefore, their reviews rarely include comments regarding the validity of statistical procedures. On the other hand, some statistically trained reviewers differ regarding the validity of the statistical analysis presented in the same manuscript. Thus, it is sometimes problematic to decide which statistical recommendation to make to the authors in order for them to refine their conclusions from their data.

To reinforce the review of the statistical analyses and provide authors with the best recommendation, since June 2021 we have included a new editorial role: the *statistical editor*. The purpose is to strengthen our editorial team from the statistical and methodological point of view. The principal function of the statistical editor will be to verify whether the statistical procedures presented by the authors are valid in light of their sampling design and the nature of the information. The statistical editor's review will be before or simultaneous with the evaluation of the peer reviewers so that the editor-in-chief will have a more informed assessment. In addition, the statistical editor will be in-charge of the statistical literacy section that we will incorporate in future journal issues. All the above will ultimately benefit the authors as it will prevent their manuscripts from being weakened by poor statistical analysis.

Given the current crisis of experimental replication, the loss of confidence in the frequentist paradigm, and the "statistical war" between frequentists and Bayesians (Gelman & Loken 2014, Wasserstein & Lazar 2016, Amrhein *et al*. 2017), professional statistical advice is more necessary than ever. In addition, many authors fall into serious statistical "sins" due to an obsession with significant findings (perhaps in pursuit of fame) or only to publish more papers faster and therefore to appear highly productive (Comroy 2019a, Makin & Orban de Xivry 2019, Orban de Xivry in Comroy 2019b). Thus, we consider it imperative that authors be forewarned of these situations, so that they might wisely decide how to approach their statistical analyses. It is also essential that authors, reviewers, and editors place the philosophical principles of statistics above statistical algorithms. As ornithological journals' editors, we are in a keystone position to promote changes that will lead us to use statistics wisely. That requires that authors and reviewers be clear about the fundamental meaning of statistics and the strengths and weaknesses of existing statistical paradigms (Underhill 1999, Colling & Szűcs 2021).

How do ornithologists use statistics?

Ornithologists use statistics both descriptively and inferentially. In the first case, we analyze and summarize information using tables or graphs of frequency distribution, measures of central tendency or location (*e.g*., mean, median, mode), and variability (*e.g*., variance, standard deviation). Such statistical descriptors are valuable for establishing, for example, differences in body size (*e.g*., Egli 1996) or trends in species richness or abundance. Although descriptive statistical procedures do not require theoretical assumptions and do not necessarily depend on experimental designs, they are essential for moving toward inferential statistics.

In the case of inferential statistics, it is fundamental to decide whether our hypotheses' predictions are or aren't admissible (Gotelli & Ellison 2004, León-Guerrero & Frankfort-Nachmias 2018). Inferential procedures are strongly conditional to the rigorous use of the scientific method. Before using any inferential method, we must establish an explicit research question, and one or several hypotheses, articulate its predictions, design and execute experiments, and collect, organize, and summarize the information generated by our experiments (Quinn & Keough 2002, Gotelli & Ellison 2004). Using inferential statistics, we draw conclusions about large groups (populations) from small groups (samples). When we do this, we presume that samples represent the population, are randomly selected, and that our data are reliable.

Currently, ornithologists follow three inferential statistical paradigms: classical or frequentist statistics, Bayesian statistics, and model selection based on information theory. The frequentist paradigm relies on the classical concepts of probability, hypothesis testing (null [H0] vs alternative hypothesis [H1]), significance testing, the significance level or alpha threshold (α), and probability value (p-value) (Fisher 1935). Within a defined study framework, researchers reject or fail to refuse a hypothesis based on a statistical significance level (Fisher 1935). The term frequentist alludes to the long-term frequency of results in infinite repetitions of experiments or samples (Johnson 1999).

Ornithologists who follow Bayesian statistical procedures measure the effect of a treatment by incorporating prior information about a given event to calculate its posterior probability. They quantify the evidence strength based on the Bayes factor. The higher the Bayes factor, the higher the evidence strength. Ornithologists who follow this line pre-determine that probability either subjectively (*e.g*., "founded belief") or objectively (*e.g*., previous measurements) (van de Schoot *et al*. 2021). Ornithologists driven by the information-theoretic approach select, among several candidate models, those with the variables with the higher explanatory power using selection criteria that minimize information loss (Burnham & Anderson 2001). Since each variable in a model represents a hypothesis, model selection is consistent with the concept of multiple hypotheses (Chamberlin 1890).

Among the statistical paradigms described, frequentist statistics is prevalent in ornithology. The reason is that most scientists and academics have accepted, used, and promoted this paradigm for more than five decades. Until a few years ago, the frequentist approach was the only one taught in universities and widely explained in dozens of books. Therefore, many generations of ornithologists graduate from their careers with an "essentially frequentist mind." Although the frequentist paradigm has several virtues, it suffers from a crisis of confidence due to questions about the inferential utility of statistical significance tests. However, the causes of this crisis resulted mainly from the misuse and abuse of significance tests and the misinterpretation of the p-value. This is partly a consequence of the poor teaching of statistics in universities.

Here we provide suggestions that will guide authors on how to wisely approach the use of inferential statistics before submitting their manuscripts to our journal. Although we focus primarily on the frequentist paradigm, many of our suggestions will enlighten those looking at other statistical paradigms. We know many authors are well-familiar with the statistical resources necessary for their research. However, other authors have only an incipient notion of the use of statistics. Thus, our suggestions are directed mainly to those authors with little or no experience with inferential statistics and those who use it only as a ritual within the scope of their research. In materializing our suggestions, we followed the reflective line of Peter Feinsinger (Feinsinger 2001) and the foundations of statistical inference according to Ronald Fisher (Fisher 1922, 1956). We enriched these suggestions with several reflections from other authors.

At the end of this editorial, we include a glossary of several of the technical and philosophical terms we use throughout our story. Although these terms are elementary, we do not always understand them so clearly, and some philosophical concepts are not easy to define and assimilate satisfactorily.

Dear author, here are our suggestions:

1. Appropriate use of any inferential statistical procedure requires that your results come from a study based on a robust experimental design. Ensure that your study units are independent (*i.e.*, that neither influences nor affects the observations recorded in the other), are randomly distributed, and are randomly selected (James & McCulloch 1985, Kamil 1988). In many field situations, it is not possible to guarantee the condition of independence or to randomly assign study units. Therefore, you should be cautious of your conclusions and state any design limitations in your manuscript (see suggestions 20). Ideally, you should obtain an appropriate sample size to satisfactorily answer your research question or to get support for your biological hypothesis (see Recommendation 3). The sample size calculation should consider, as a minimum, some empirical criteria (*e.g.*, ≥ 20 replicates; Simmons *et al*. 2013). Without these essential requirements, the use of any inferential statistics will lead to spurious or distorted results. There is a varied supply of books on experimental design. For the design of field studies, we recommend the books by Feinsinger (2001) and Scheiner & Gurevitch (2001) and the chapter on the design of experiments in the book by Gotelli & Ellison (2004).

2. When you cannot obtain a statistically adequate number of study units, you can follow "the Rule of 10" (Gotelli & Ellison 2004). This rule has no theoretical basis but reflects the experience gained in the field with successful and unsuccessful experimental designs. In exceptional cases (*e.g.*, a naturally rare species), a minimum of five sampling units (*e.g.*, nest site) per condition (*e.g.*, habitat type) may be acceptable. *Ad hoc* statistical tests exist for these exceptions. However, your statistical power will be lowest (see suggestion 3), and you run the risk that your inferences will be very limited.

3. Consider measuring statistical power before designing your study (see Glossary). When calculating the sample size, you should estimate how big is the type II error (see Glossary). Statistical power tells you whether the number of study units you determined was sufficient to detect an effect when that effect exists (Steidl & Thomas 2001, Lecoutre *et al*. 2005). The higher the statistical power, the greater the chance of detecting a true effect. However, be aware that high statistical power does not always guarantee the finding of a true effect (Button *et al*. 2013). We also recommend calculating statistical power after you have gotten the sample so that you are able to realize the inferential limitations of your study. Overall, field studies achieve low to medium statistical power due to the difficulty of obtaining sufficient replicates (Halsey *et al*. 2015, Parker *et al*. 2018). Knowing the statistical power of

your study will also guide you regarding the robustness of your experimental design to the possibility of using non-frequentist statistical methods (Steidl & Thomas 2001).

4. On many occasions, you may have difficulty obtaining a statistically appropriate number of study units (*e.g.*, < 10) and will have to settle for very few (see recommendation 3). In these cases, it is practical to visualize your data graphically and qualitatively identify any trends or associations (Ellison 2001). You can also describe your data using measures of central tendency (*e.g.*, mean, median, mode) or dispersion (*e.g.*, range, standard deviation). There are several frequentist statistical tests designed to analyze small sample sizes (Zar 1999). However, before you decide to use them, check whether they are valid for studies based on observational experiments.

5. Plan your statistical analysis before implementing your study and use the most straightforward inferential procedure that fits your experimental design. If your results come from field studies, you should prefer nonparametric tests or models. Overall, the nonparametric methods are appropriate for small sample sizes and have fewer restrictions than their parametric analogs (Fowler & Cohen 1994, Siegel & Castellan 1995). Nonparametric procedures only require we sort the different measures of a variable of interest that we wish to compare across data sets.

6. Always have at hand appropriate statistical texts for your study field. In the case of simple experimental designs, we recommend the book by Fowler & Cohen (1994). A quite didactic and entertaining book is Salkind (2017). Always ensure that the texts contain descriptions of a wide range of nonparametric statistical tests. We recommend books by Zar (1999) and Siegel & Castellan (1995). Sokal & Rohlf (1969) provide quite didactic descriptions of some nonparametric procedures. Nonparametric procedures are often the most appropriate for the analysis of results from field studies (see Recommendation 10). All the above books have a pure frequentist approach. A text that also contains descriptions of non-frequentist procedures is Quinn & Keough (2002). Two good books to open the Bayesian mind are by McElreath (2016) and Field (2016). Most of these books are in English, but several basic statistics books in Spanish are available on-internet (*e.g.*, Blair & Taylor 2008).

7. Always reflect on whether you need to apply any inferential statistical procedure. Even if your experimental design is sound and your sample is large, statistical tests or models are often unnecessary or inappropriate for analyzing your results (Sokal & Rohlf 1969, Cherry 1998, Johnson 1999). Many field experiments are subject to environmental variables not considered by the researcher. If you do not control such variables, it is best not to apply any inferential procedures. In addition, for logistical reasons or natural limitations (*e.g.*, natural scarcity of a habitat type), you may not be able to meet the assumption of spatial independence of your study units. In these cases, you could obtain an enormous amount of interesting biological information whose value will not be increased by an inferential statistical analysis. Indeed, that analysis would be inappropriate. There is more. Your experimental design may be indisputably robust, but if the effect size is evident, you don't need to "garnish" your finding with any statistical algorithms. Again, a graphical display of your results in combination with descriptive statistical resources may be the best way to extract the best information from your data (Cherry 1998, Lang *et al*. 1998, Ellison 2001, Cumming 2012).

8. Before using procedures based on statistical significance tests, be aware that such tests possess several limitations that make them inappropriate or useless for observational studies (Buckland 1982, Cherry 1998, Johnson 1999, Greenland & Poole 2013). Overall, statistical significance tests are uninformative and logically poor, the α level is arbitrary and without theoretical basis, and the p-value is vulnerable to misinterpretation (*e.g.*, Yoccoz 1991, Anderson *et al*. 2000, Martínez-Abraín & Oro 2005, Greenland & Poole 2013). Furthermore, the p-value leads to dichotomous decisions, such as to admit or not admit the null hypothesis and doesn't allow us to quantify the direction of the difference found. Finally, the p-value is sensitive to sample size (Underhill 1999). For example, if two ornithologists conduct the same experiment with different sample sizes, say ten vs 30, they might reach opposite conclusions (*i.e.*, $p > 0.05$ vs $p < 0.05$).

9. When justifiably using tests of statistical significance, do not label your findings using the terms "statistically significant" (*e.g.*, $p < 0.05$) or "statistically non-significant" (*e.g.*, $p > 0.05$). In common parlance, persons interpret the term "significant" as reliable. However, a p-value < 0.05 and a p-value > 0.05 are not necessarily evidence that the null hypothesis is false or true, respectively (Amrhein *et al*. 2019). Remember that α is an arbitrary value!

10. Avoid the "significant-itis" (*i*.e., the compulsive obsession with statistically significant findings; Chia 1997). Many authors living in the "p-value culture" (Nelder 1999) pursue statistically significant results because (i) they confuse the p-value as an indicator of the evidence strength, and (ii) it is the only way for their manuscripts to be acceptable to editors and reviewers with the same confusion (Underhill 1999, Martínez-Abraín & Oro 2005, Sedgwick 2022). The "p-value culture" may even lead some researchers to p-value hacking (*i.e*., pushing their p-values toward significant limits to make their findings appear relevant and publishable [Comroy 2019a]). However, statistical significance does not necessarily reflect the biological or clinical relevance of our findings (Potish *et al*. 1980, Krebs 1989, Yoccoz 1991, Underhill 1999, Malay 2016). The statistically little differences can have biologically considerable consequences. It is always worthwhile to remember that Fisher's intention was that we use statistical significance only as a tool to indicate that our results warrant further investigation (Fisher 1935, Sedgwick *et al*. 2022).

11. When appropriate, calculate confidence intervals around the measure of interest (*e.g*., the mean) for different levels or groups of samples (Martínez-Abraín & Oro 2005). Confidence intervals let you know how reasonable your estimates are, reliably compare estimates from distinct sample groups, and whether there are any biologically significant effects (*i.e*., the effect size). In most studies, estimation will be more important than applying a statistical test (Yoccoz 1991, Cherry 1998, Krebs 2000). Always graphically compare your confidence intervals to visualize the magnitude of the effect. The latter increases the informative power of your analyses (Cherry 1998, Johnson 1999, Ellison 2001).

12. Consult a professional biostatistician when you do not understand or have doubts about using a complex statistical procedure (Buckland 1980, Gustavii 2008). If your study is observational, look for a biostatistician who has experience with observational designs and knows the limitations of information from studies conducted in the field. Be aware that some biostatisticians may disagree about using some statistical procedures, which may confuse you further. For example, one professional may require you to use statistical modeling to analyze your results from just five sampling sites. However, another may suggest you only evaluate your data with descriptive statistics. In these cases, be wise and opt for the most straightforward and informative solution.

13. Do not use sophisticated statistical procedures only to surprise reviewers or show that your study is statistically advanced, especially if you have few observations or if your study is merely observational. Because multivariate analyses "process" multiple questions or multiple hypotheses at once, researchers must meet several theoretical and empirical assumptions. In addition, users of these analyses must perform several non-explicit tests and adjustments before inferring anything from their information. The latter leads to a critical problem. With each additional test, the probability increases that a researcher will erroneously conclude that there is at least one "statistically significant" effect (Gelman *et al*. 2012, Gelman & Loken 2014). When researchers use multivariate procedures to draw inferences from exploratory or observational studies, their conclusions are often spurious (James & McCulloch 1990).

14. When you justifiably use sophisticated statistical procedures, try to describe them as clearly, readily, didactically, and intuitively as possible (*e.g*., Fowler & Cohen 1996, Gustavii 2008, In & Lee 2017). When reporting your results, do so considering readers' "left brain" and "right brain" in mind, *i.e*., do not just deliver "hard" numerical values to support your conclusions, but also enlighten your readers by pleasantly relating the extent of those values (Buckland 1980). Remember that our journal focuses on a broad audience. Therefore, the universe of our readers may have a widely varying level of statistical knowledge. Anderson *et al*. (2001) and Brennan (2012) offer guidelines for informatively presenting the results of statistical analyses within a manuscript. Thirty-six years ago, John Gerrard (1985) stated that *the interplay between amateur and professional workers is one of the strengths of ornithology compared to other areas of scientific investigation*. However, Gerrard also warned that this strength was at risk because professional ornithologists were using sophisticated statistical techniques incomprehensible to amateur ornithologists. No doubt that caution still stands.

15. Do not assume that statistical programs, no matter how sophisticated, provide truly reliable results (*e.g*., Eklund *et al*. 2016). Computer programs allow us to save time and minimize hand work, but they are fallible (Littlewood & Strigin 1992). Moreover, statistical programs do not discriminate whether your data come from a robust or flawed experimental design. On the

other hand, programs designed for model simulation can generate exacerbated values of a statistical index and, thereby, could lead you to spurious conclusions (White *et al*. 2014). If your results seem "suspects" or biologically meaningless, rethink your analyses or seek expert guidance. Always be willing to reconsider your data and analysis results (Carraway 2009). It will always be revealing to contrast the statistical indicators provided by your computer program with a graphical representation of your results. If you find that they are congruent, then rest assured. If not (again), cultivate your wise and opt for the more straightforward yet informative alternative.

16. Be wary of "hype" about the superior virtues of emerging statistical paradigms. Several researchers have made an intense and widespread call in the literature to replace frequentist statistics with Bayesian statistics or model selection (*e.g*., Anderson *et al*. 2000, Anderson & Burnham 2002, Ellison 2004). Although these same researchers call for the appropriate use of these emerging statistical paradigms, many others tend to apply them inappropriately. If you wish to look into these emerging paradigms, we recommend that you become well-informed about their theoretical and methodological bases. In such cases, it is strictly necessary you seek the advice of a specialized biostatistician.

17. Don't despise philosophy. We use statistics to decide what is true and what is not (Goodman 2016, Sedgwick 2022). Therefore, we need a basic understanding of the philosophical underpinnings of each statistical paradigm to informally decide which one we will follow (Sedgwick 2022). Surprising as it may seem, many ecologists, including ornithologists, are unaware of the paradigms within which they operate (Krebs 2000). Knowing the philosophical basis of each statistical paradigm will also guide you about the scope of your inferences. Statistics help us in the search for the truth contained in nature. However, statistical paradigms are human constructs and, therefore, are fallible (Sedgwick 2022). Thus, we run the risk of such constructs leading us down a path away from the truth (Greenland & Poole 2013). It may be uncomfortable to put down your computer keyboard and no longer "run" your "flagship" statistical program. It may be boring for you to read about statistical philosophy. However, you don't need to read all about Popper or Bayes; a good encyclopedia of statistical philosophy will suffice (*e.g*., Zalta & Nodelman 2021). If you can get a little notion about the philosophical foundations of inferential statistics, you will resist the "siren calls" of the emerging statistical paradigms.

18. Try to educate yourself in statistics and learn what is substantial in the field. We know that university courses in statistics strongly incline toward the mechanical use of statistical algorithms. Instructors accentuate that practice by using the various offers of computer programs. Even worse are statistics courses in non-statistical postgraduate studies. Generally, instructors of these "advanced" courses overwhelm students from week to week with dozens of statistical articles. That is what Salkind (2017) aptly calls "sadistics". The absence of pedagogy degrades the quality of statistics courses (Zieffler 2018). The good thing is that there are many resources on the internet to learn statistics in a friendly, effective, and autonomous way. The portal of www.youtube.com is abundant in tutorials to introduce statistical concepts and procedures. Of course, some wise authors write practical and engaging books on statistics (*e.g*., Fowler & Cohen 1994, Field 2016, Salkind 2017). One rather fun book focused on giving meaning and "flavor" to statistics is Huff (1993).

19. Krebs (2000) wisely advised us to spend more time on real ecological issues. For better or worse, inferential statistics cuts across many matters in ecology. When you are in the field, reflect on the rationale for statistical inference in ornithology. Remember that statistical inference is the process of drawing conclusions about a population using information from a set of (often small) samples from that same population. When you visit your sampling units in the field, you will notice that you are unlikely to get each sample under the same conditions. That is because each sampling unit will vary spatially and temporally and because your perception may vary in each sampling unit. The latter is difficult to control statistically. However, the prominent virtue of statistics is that it helps us deal with the uncertainty and variability inherent in the natural world (Sokal & Rohlf 1969). Thus, rather than viewing statistical inference as a process of reaching conclusions, we should see it as a way of accepting and measuring the uncertainty and variability in our information (Mallows 1998, Gelman 2016, 2019, Amrhein *et al*. 2017, Tong 2019).

20. Always acknowledge explicitly in your manuscript the limitations of your experimental designs and statistical analyses. When aware of your statistical errors, admit them and correct them promptly (Gelman

2018, 2020). Do control speculation. Do not conclude or infer anything beyond what your results reveal. Generally, extrapolations are risky in a statistical context (Buckland 1980). Honesty and transparency are fundamental to the integrity of science (Gelman 2017, 2018, Parker *et al*. 2018). Authors who report appropriately are more credible. Often our biases and ego make us forget that we are fallible and that our studies are not perfect (Gigerenzer 1993, Lanni 2021). Admitting our mistakes is the essence of scientific progress (Kareiva & Marvier 2018). Vuilleumier (2004) made us realize that learning from mistakes allows tremendous theoretical advances in ornithology.

21. Keep the following messages in mind. Statistics is not a toolbox for research purposes. The true meaning of statistics is to learn from information and to measure, control, and communicate the uncertainty and variability contained in that information (Wild *et al*. 2018). Statistical procedures applied judiciously and modestly protect us from false findings (Cox 2001). The wise and honest use of statistics helps us understand the surrounding world and bring us closer to the truth we seek within it (Salkind 2017, Sedgwick 2022). Statistics can help us see the world clearly if we are willing to look (Tarran 2020).

Our commitment as editors

The judicious use of statistics helps to reveal the links between our findings and the theories in which we frame our studies. The results of statistical analysis permit us to evaluate our hypotheses and theories, uncover unexpected patterns and trends, and provide the impetus to reformulate the approaches we work with (Krebs 2000, León-Guerrero & Frankfort-Nachmias 2018). However, a statistical procedure is only a tiny part of the research process, and even the most rigorous statistical method will not reveal the whole truth we are pursuing.

On the other hand, all scientific research is vulnerable to statistical misuse and researcher bias (Gelman 2018, Kareiva & Marvier 2018). The (often unintentional) misuse and abuse of statistics combined with poor experimental designs have resulted in many spurious or at least questionable findings (Ioannidis 2005, Ioannidis *et al*. 2014). That can considerably delay progress in ornithology (Martínez-Abraín & Oro 2005). Your duty as an author is to always match your statistical analyses appropriately and realistically to your experimental design and scientific reasoning (Goodman 2016). Like Buckland (1980), we hope our suggestions will encourage ornithologists to use statistics skillfully and wisely and reorient those who apply them inappropriately.

Both editors and reviewers of scientific journals are the arbiters of scientific practice (Cherry 1998, Johnson 1999, Parker *et al*. 2018, Sedgwick *et al*. 2022). Therefore, we do not only need to understand how statistical procedures work but to convey their appropriate use to authors (Johnson 1999, Underhill 1999). The latter includes cautioning authors to frame their conclusions within the limitations of their study. We should have no objections when the authors present well-meshed statistical analyses with a sound experimental design. However, if the authors present poor or inadequate statistical analyses, our ethical and professional responsibility is to suggest to them the wisest options (Parker *et al*. 2018). The latter might even include discarding any inferential statistical analysis. In the face of poor statistical analysis, we editors and reviewers have a responsibility to suggest changes that appropriately ensure both the credibility of the findings and the authors' prestige (Ioannidis 2014). Although our role is to "judge" the scientific merit of each manuscript, one of our premises is not to reject a manuscript due to analytical errors that are correctable. As editors of the *Revista Chilena de Ornitología*, we will always be available for statistical assistance to authors who require it.

GLOSSARY...OF SOME STATISTICALLY SIGNIFICANT TERMS!

Alpha (α), α level, α value, "significance level": is the probability of making a type I error (see below). A researcher must adopt or decide which α to use before the design of experiments. We can interpret α as a numerical value set "nominally" based on experience and that we expect to make an error. Ornithologists typically adopt an $\alpha = 5\%$.

Experimental design: any planned experiment, whether it is a well-controlled experiment (typically a laboratory experiment), a manipulative experiment in the field (*e.g.*, natural precondition vs intervened postcondition), or an observational or mensurative (*i.e.*, comparison of two or more natural situations without the intervention of the researcher) (James & McCulloch 1985, Scheiner & Gurevitch 2001, Gotelli & Ellison 2004).

Type I error: we commit a type I error when we reject the null hypothesis as being true.

Type II error: we commit a type II error when we do not reject the null hypothesis as being false.

Evidence: fact or a piece of information that supports what we believe, postulate, or hypothesize (Zalta & Nodelman 2021).

Hypothesis: is a general proposition that suggests an explanation for an observed phenomenon (Krebs 2000). We can verify only indirectly one hypothesis by examining its predictions (Farji-Brener 2004).

Alternative hypothesis (H_1): is a hypothesis of difference or association between treatments or experimental conditions: *i.e.*, that suggests an effect. It's the hypothesis that the researcher postulates as valid based on the patterns observed in his or her data, which differ from what the null postulates.

Null or null hypothesis (H_0): is a hypothesis of no difference or association between treatments or experimental conditions. Under significance testing, this hypothesis is admissible until there is no evidence to the contrary and the stated value of the parameter is valid. A researcher relies on a null to explain the patterns observed in the data in a simple way, attributing any variation in the data to chance or measurement error (Gotelli & Ellison 2004).

H_0 and H_1 are essentially statistical hypotheses; that is, they are restricted, operational guesses about the value of a particular (population) "parameter" and represent expected outcomes under different biological scenarios (Farji-Brener 2004).

Statistical inference: the process of drawing conclusions about the population using information from a sample or a set of samples from that population. Statistical inference is made possible by mathematical methods.

Confidence interval: is the range of values around a statistical estimator's value (*e.g.*, mean estimate) obtained from a sample and which theoretically will include with a high probability the value of the population parameter. Usually, ornithologists estimate 95% confidence intervals (Fowler & Cohen 1994). That is a confidence interval that includes, for example, the population mean in 95 trials out of 100. The advantage of confidence intervals is that they allow a quantitative assessment of the magnitude of the effect and its precision.

Paradigm: is a worldview; a broad approach to the problems addressed in a field of science (Krebs 2000).

Statistical parameter: is the value of a mathematical function that summarizes the information about a variable under study (*e.g.*, mean, variance, standard deviation). The term parameter came from the Greek parametron meaning beyond measurement. Since we do not know the value of the population parameters, we calculate that value by means of the measurements obtained in a sample or several samples of that population by means of a mathematical function called "estimator" while "estimation" is the specific calculation related to one sample. Thus, we should not confuse "estimation" with "parameter"; a statistical parameter will always be "beyond our measurement."

Statistical power: is a measure of the ability of an experiment to detect an effect when that effect exists (Button *et al*. 2013, Halsey *et al*. 2015). Mathematically, statistical power is the complement of beta: $1 - \beta$. Thus, the higher the statistical power, the lower the probability of making a type II error and the higher the chance of detecting an effect. Although the level of statistical power depends on several factors, the sample size is decisive. If an ornithologist bases an experimental study on a small sample size, it will have low statistical power. Depending on the study type, a researcher may exaggerate the effect magnitude when using samples too small. For this reason, when one expects effect sizes to be weak or moderate, it is advisable to avoid inferences based on the p-value threshold when the sample size is small (*e.g.*, < 10).

Prediction: predictions are expected outcomes assuming that our research hypothesis is true. That is to say, the null and alternative express precisely the result we expected if our biological hypothesis is correct (Farji-Brener 2004).

Statistical significance: there are many ways to think about and approach the concept of statistical significance. Usually, we call statistical significance the α value. That is relevant when we perform hypothesis testing, since the rejection or not of the null hypothesis depends on whether the p-value remains under or over α. We give the most conservative definition based on the Fisherian paradigm and which considers statistical reasoning beyond the analytical contrast:

We reach statistical significance in a sample under study when the results are observable in the population on which we perform the statistical sampling (probabilistic) and are not due to chance, but these respond to patterns of evidence consistent enough to be observable in samples coming from that population.

Due to the impossibility of studying the population, we try to observe this pattern in the sample (usually a single sample) and infer what happens in the population by means of statistical methods of contrast based on statistical significance tests. The statistical significance testing permit us to rule out chance as the explanation for what we observed. For this reason, we randomized the samples. We performed a randomization process to avoid systematic deviations induced by the experimenter when selecting the units under study.

Effect size: is the magnitude of an effect observed after an experimental treatment. Depending on the question, a researcher measures effect size by quantifying the difference between two treatments or conditions or measuring the association strength between two variables or processes (Button *et al*. 2013). Often, researchers use standardized categories of effect size (*e.g*., weak, moderate, strong; see Cohen 1988). The effect measured in a sample is an estimate of the "true" effect size in the population. Usually, we interpret the p-value by assuming the true effect size is zero (Halsey *et al*. 2015).

p-value: is conceptually the probability that a statistical estimator (*e.g*., correlation coefficient) is as or more extreme than its observed value (estimate) given the null hypothesis (Wasserstein & Lazar 2016). Put another way, the p-value is the probability of making a type I error if we reject the null hypothesis based on the data analysis. After comparing the p-value with α, we adopt the rule that if $p < \alpha$, we reject H_0 (*i.e*., retain H_1), and if $p > \alpha$, we admit H_0 (*i.e*., reject H_1). The equality of p and α is controversial, and we will address it later in the statistical literacy section.

Truth: is a profoundly philosophical concept, and its definition (if it is definable at all) is subject to various ideologies (Zalta & Nodelman 2021). In scientific terms, we can mean "truth" as the set of invariant or undeniable properties of nature (*e.g*., organisms, patterns, processes, mechanisms) (Cohen 1980). At least, this is the truth that interests us as ornithologists. Many of nature's properties are invisible or intangible, and we will look for evidence that brings us closer to them. In doing so, we postulate hypotheses, design experiments, analyze our results, and apply inferential procedures. If everything is well, our inferences will reflect something of the truth we seek.

ACKNOWLEDGMENTS. - We thank Dr. Karla Yohannessen, School of Public Health, Universidad de Chile, for reviewing an earlier version of this editorial. Her comments and suggestions allowed us to refine the text.

LITERATURE CITED

Amrhein, A., S. Greenland & B. McShane. 2019. Retire statistical significance. *Nature* 567: 305-307.

Amrhein, V., F. Korner-Nievergelt & T. Roth. 2017. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ* 5: e3544. https://peerj.com/articles/3544

Anderson, D.R. & K.R. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management* 66: 912-918.

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

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

Blair, R.C. & R. Taylor. 2008. *Bioestadística*. Pearson Educación. Naucalpan de Juárez, México. 552 pp.

Brennan, L. 2012. Editorial guidance and wildlife science: the roles of Wildlife Society Bulletin authors, associate editors, and reviewers. *Wildlife Society Bulletin* 36: 392-398.

Buckland, S.T. 1982. Statistics in ornithology. *Ibis* 124: 61-66.

Burnham, K.P. & D.R. Anderson. 2001. Kullback–Leibler information as a basis for strong inference in ecological studies. *Wildlife Research* 28: 111-119.

Button, K.S., J.P.A. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S.J. Robinson & M.R. Munafo. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14: 365-376.

Carraway, L.N. 2009. Ethics for and responsibilities of authors, reviewers and editors in science. *American Midland Naturalists* 161:146-164.

Chamberlin. T.C. 1890. The method of multiple working hypotheses. *Science* 148: 754-759.

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

Chia, K-S. 1997. "Significant-itis" – an obsession with the P-value. *Scandinavian Journal of Work Environmental Health* 23: 152-154.

Cohen, J. 1980. What has science to do with truth? *Synthese* 45: 489-510.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Second edition. Lawrence Erlbaum Associates, New York, U.S.A. 567 pp.

Colling, L.J. & D. Szűcs. 2021. Statistical inference and the replication crisis. *Review of Philosophy and Psychology* 12: 121-147.

Comroy, G. 2019a. The 7 deadly sins of research. *Nature News*. https://www.nature.com/nature-index/news-blog/the-seven-deadly-sins-of-research

Comroy, G. 2019b. The 10 most common mistakes with statistics, and how to avoid them. *Nature News*. https://www.nature.com/nature-index/news-blog/ten-most-common-mistakes-with-statistics-and-how-to-avoid-them

Cox, D.R. 2001. Another comment on the role of statistical methods. *Physical Therapy* 81: 1469.

Cumming, G. 2012. *Understanding the new statistics - effect sizes, confidence intervals, and meta-analysis.*

Routledge Taylor & Francis Group, New York, U.S.A. 519 pp.

Egli, G. 1996. Biomorfología de algunas aves de Chile central. *Boletín Chileno de Ornitología* 3: 2-9.

Eklund, A., T.E. Nichols & H. Knutsson. 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of National Academy of Sciences* (*PNAS*) 113: 7900-7905.

Ellison, A.M. 2001. Exploratory data analysis and graphic display. Pp. 37-62, *in* Scheiner, S.M. & J. Gurevitch (eds.) *Design and analysis of ecological experiments*. Oxford University Press, New York, U.S.A.

Ellison, A.M. 2004. Bayesian inference in ecology. *Ecology Letters* 7: 509-520.

Feinsinger, P. 2001. *Designing field studies for biodiversity conservation*. The Nature Conservancy. Island Press, Washington, U.S.A. 212 pp.

Farji-Brener, A. 2004. ¿Son hipótesis las hipótesis estadísticas? *Ecología Austral* 14: 201-203.

Field, A. 2016. *An adventure in statistics - the reality enigma*. SAGE Publications, London, U.K. 746 pp.

Fisher, R. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London (Series A)* 222: 594-604.

Fisher, R. 1935. *The design of experiments*. Oliver and Boyd, Edinburgh, U.K. 257 pp.

Fisher, R. 1956. Statistical methods and scientific induction. *Journal of the Royal Statistical Society* (Serie B) 17: 69-78.

Fowler, J. & L. Cohen. 1994. *Statistic for ornithologists*. Second edition. British Trust for Ornithology, London, U.K. 150 pp.

Gelman, A. 2016. The problems with *P*-values are not just with *P*-values. Supplemental material to the ASA statement on *P*-values and statistical significance. *American Statistician* 70: 129-133.

Gelman, A. 2017. Honesty and transparency are not enough. *Chance* 30: 37-39.

Gelman, A. 2018. Ethics in statistical practice and communication: five recommendations. *Significance* 15: 40-43.

Gelman, A. 2019. The principles of uncertainty. *Nature* 569: 628-629.

Gelman, A. 2020. Evidence vs. truth. http://www.stat.columbia.edu/~gelman/research/published/ethics22.pdf

Gelman, A. & E. Loken. 2014. The statistical crisis in science. *American Scientists* 102: 460-465.

Gelman, A., J. Hill, & M. Yajima. 2012. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5: 189-211.

Gerrard, J. 1985. Statistical analysis in ornithology. *Ringing & Migration* 6: 60.

Gigerenzer, G. 1993. The superego, the ego, and id in statistical reasoning. Pp. 311-339, *in* Keren, G. & C. Lewis (eds.) *A handbook for data analysis in the behavioral sciences: methodological issues*. Lawrence Erlbaum Associates, New Jersey, U.S.A.

Goodman, S.N. 2016. Aligning statistical and scientific reasoning - misunderstanding and misuse of statistical significance impede science. *Science* 352: 1180-1181.

Gotelli, N.J. & A.M. Ellison. 2004. *A primer of ecological statistics*. Sinauer Associates Inc., Sunderland, Massachusetts, U.S.A. 510 pp.

Greenland, S. & C. Poole. 2013. Living with statistics in observational research. *Epidemiology* 24: 73-78.

Gustavii, B. 2008. *How to write and illustrate scientific papers*. Second edition. Cambridge University Press. Cambridge, U.K. 168 pp.

Halsey, L.G., D. Curran-Everett, S.L. Vowler & G.B. Drummond. 2015. The fickle *P* value generates irreproducible results. *Nature Methods* 12:179-185.

Hampel, F. 1997. *Is statistic too difficult*? Research Report N° 81. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland. 18 pp.

Huff, D. 1993. *How to lie with statistics*. W.W. Norton & Company Inc., New York, U.S.A.

In, J. & S. Lee. 2017. Statistical data presentation. *Korean Journal of Anesthesiology* 70: 267-287.

Ioannidis, J.P.A. 2005. Why most published research findings are false. *Public Library of Science (PLoS) Medicine* 2: e124.

Ioannidis, J.P.A. 2014. How to make more published research true. *Public Library of Science (PLoS) Medicine* 10: e1001747.

Ioannidis, J.P.A., S. Greenland, M.A. Hlatky, M.J. Khoury, M.R. Macleod, D. Moher, K.F. Schulz & R. Tibshirani. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 383: 166-175.

James, F.C. & C.E. McCulloch. 1985. Data analysis and the design of experiments in ornithology. Pp. 1-63, *in* Johnston, R.F. (ed.) *Current Ornithology*. Plenum Press, New York, U.S.A.

James, F.C. & C.E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or pandora's box? *Annual Review of Ecology and Systematic* 21:129-166.

Johnson, D.H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management*

63: 763-772.

Kamil, A.C. 1988. Experimental design in ornithology. Pp. 313-346, *in* Johnston, R.F. (ed.) *Current Ornithology*. Plenum Press, New York, U.S.A.

Kareiva, P. & M. Marvier. 2018. Uncomfortable questions and inconvenient data in conservation science. Pp. 3-9, *in* Kareiva, P., M. Marvier & B. Silliman (eds.) *Effective conservation science: data not dogma*. Oxford University Press, Oxford, U.K. 190 pp.

Krebs, C.J. 1989. *Ecological methodology*. Harper Collins Publisher. New York, U.S.A. 654 pp.

Krebs, C.J. 2000. Hypothesis testing in ecology. Pp. 5-14, *in* Boitani, L. & T. Fuller (eds.) *Research techniques in animal ecology: controversies and consequences*. Columbia University Press, New York, U.S.A. 442 pp.

Lang, J.M., K.J. Rothman & C.L. Cann. 1998. That confounded *P*-value. *Epidemiology* 9: 7-8.

Lanni, J. 2021. A lesson of failure. *Science* 374: 1642.

Lecoutre, B., J. Poitevineau & M-P. Lecoutre. 2005. A reason why not to ban null hypothesis significance tests. *Revue MODULAD* 33: 249-253.

León-Guerrero, A. & C. Frankfort-Nachmias. 2018. Essentials of social statistics for a diverse society. Third edition. SAGE Publications, Thousand Oaks, California, U.S.A. 547 pp.

Littlewood, B. & L. Strigin. 1992. The risks of software. *Scientific American* 267: 62-75.

Malay, D.S. 2016. Editorial: clinical significance, clarified. *Journal of Foot & Ankle Surgery* 55: 903-905.

Mallows, C.L. 1998. The zeroth problem. *American Statistician* 52: 1-9.

Martínez-Abraín, A. & D. Oro. 2005. Can ornithology advance as a science relying on significance testing? A literature review in search of a consensus. *Ardeola* 52: 377-387.

McElreath, R. 2016. *Statistical rethinking - a Bayesian course with examples in R and Stan*. Chapman & Hall Book. Boca Raton, Florida, U.S.A. 469 pp.

Makin, T.R & J.J Orban de Xivry. 2019. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* 8: e48175.

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

North, P.M. & J.T. Byron. 1985. The general context of statistics in ornithology. Pp. 1-9, *in* Morgan, B.J.T & P.M. North (eds.) *Statistics in ornithology*. Springer-Verlag, Berlin, Germany.

North, P.M. 1994. Ornithological statistics. Pp. 463-506, *in* Patil, G.P. & C.R. Rao (eds.) *Handbook of statistics. Vol. 12. Environmental Statistics*. Elsevier Science Publisher. Cambridge, Massachusetts, U.S.A.

Parker, T.H., S.C. Griffith, J.L. Bronstein, F. Fidler, S. Foster, H. Fraser, W. Forstmeier, J. Gurevitch, J. Koricheva, R. Seppelt, M.W. Tingley & S. Nakagawa. 2018. Empowering peer reviewers with a checklist to improve transparency. *Nature Ecology & Evolution* 2: 929-935.

Potish, R., J. Boen & L. Seymour. 1980. The distinction between statistical significance and clinical importance. *International Journal of Radiation Oncology, Biology and Physics* 6: 941-944.

Quinn, G.P. & K.J. Keough. 2002. *Experimental design and data analysis for biologists*. Cambridge University Press, London, U.K. 553 pp.

Salkind, N.J. 2017. *Statistics for people who (think they) hate statistics: using Microsoft Excel 2016*. Fourth edition. Sage Publications, California, U.S.A. 512 pp.

Scheiner, S.M. & J. Gurevitch (eds.). 2001. *Design and analysis of ecological experiments*. Oxford University Press, New York, U.S.A. 415 pp.

Scheiner, S.M. 2001. Theories, hypotheses, and statistics. Pp. 3-13, *in* Scheiner, S.M. & J. Gurevitch (eds.) *Design and analysis of ecological experiments*. Oxford University Press, New York, U.S.A.

Sedgwick, P. 2022. Trials and tribulations of teaching null hypothesis significance testing in the health sciences. *Chance* 35: 33-41.

Sedgwick, P., A. Hammer, U.S. Kesmodel & L.H. Pedersen. 2022. Current controversies: null hypothesis significance testing. *Acta Obstetricia et Gynecologica Scandinavica* 101: 624-627.

Siegel, S. & N.J. Castellan. 1995. *Estadística no paramétrica aplicada a las ciencias de la conducta*. Fourth edition. Editorial Trillas, México D.F., México. 437 pp.

Simmons, J.P., L.D. Nelson & U. Simonsohn. 2013. *Life after P-hacking*. Meeting of the Society for Personality and Social Psychology. New Orleans, Louisiana, U.S.A. 17-19 January 2013. http://dx.doi.org/10.2139/ssrn.2205186

Sokal, R.R & F.J. Rohlf. 1969. *Biometry*. Freeman & Company, San Francisco, U.S.A. 776 pp.

Steidl, R.J. & L. Thomas. 2001. Power analysis and experimental design. Pp. 14-36, *in* Scheiner, S.M & J. Gurevitch (eds.) *Design and analysis of ecological experiments*. Oxford University Press, New York, U.S.A.

Tarran, B. 2020. “Statistics help us see the world clearly if we’re willing to look”. *Significance* 17: 42-43.

Tong, C. 2019. Statistical inference enables bad science; statistical thinking enables good science. *American Statistician* 73: 246-261.

Underhill, L.G. 1985. Avian demography: statistics and ornithology. *Ostrich* 70: 61-70.

van de Schoot, R., S. Depaoli, R. King, B. Kramer, K. Märtens, M.G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen & C. Yau. 2021. Bayesian statistics and modelling. *Nature Reviews* 1: 1. https://doi.org/10.1038/s43586-020-00001-2

Van Emden, H.F. 2008. *Statistics for terrified biologists*. Blackwell Publishing, Oxford, U.K. 343 pp.

Vuilleumier, F. 2004. A critique of Neotropical ornithology: is research on neotropical birds scientific? *Ornitología Neotropical* 15(Suppl.): 41-60.

Wasserstein, R.L. & N.A. Lazar. 2016. The ASA's statement on p-Values: context, process, and purpose. *American Statistician* 70: 129-133.

White, J.W., A. Rassweiler, J.F. Samhouri, A.C. Stier & C. White. 2014. Ecologists should not use statistical significance tests to interpret simulation model results. *Oikos* 123: 385-388.

Wild, C.J., J.M. Utts & N.J. Horton. 2018. What is statistics? Pp. 5-36, *in* Ben-Zvi, D., K. Makar & J. Garfield (eds.) *International handbook of research in statistics education*. Springer International Publishing, Switzerland. https://doi.org/10.1007/978-3-319-66195-7_1.

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.

Zalta, E.N. & U. Nodelman. 2021. *Stanford encyclopedia of philosophy*. Department of Philosophy, Stanford University, California, U.S.A. https://plato.stanford.edu/

Zar, J.H. 1999. *Biostatistical analysis*. Fourth edition. Prentice Hall, Upper Saddle River, New Jersey, U.S.A. 663 pp.

Zieffler, A. J. Garfield & E. Fry. 2018. What is statistics education? Pp. 37-70, *en* Ben-Zvi, D., K. Makar & J. Garfield (eds.) *International handbook of research in statistics education*. Springer International Publishing, Switzerland. https://doi.org/10.1007/978-3-319-66195-7_2.